Análise de retrogenes autossómicos da proteína ribossómica S4X (RPS4X) em mamíferos

Tatiana Isabel Lopes Vital
Mestrado em Genética Forense
Departamento de Biologia
2015

Orientador
Alexandra Lopes, Researcher of Population Genetics, IPATIMUP/I3S

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

Ao Professor António Amorim por ter permitido o meu ingresso neste mestrado e possibilitado a minha entrada no seu grupo de trabalho.

À minha orientadora Alexandra Lopes pela disponibilidade com que me acolheu, pela paciência com que sempre me ajudou quando eu dizia “Tenho um problema!”, pelo companheirismo, simplicidade e simpatia.

À Ana Lima por me ter possibilitado o acompanhamento e participação no seu trabalho laboratorial bem como pelos ensinamentos transmitidos.

Aos colegas da hora de almoço pela camaradagem e brincadeira que ajudaram a desanuviar das longas horas em frente ao computador.

À minha família, em especial aos meus pais por me terem possibilitado a oportunidade de investir na minha formação.

A todos os amigos que me acompanharam e apoiaram, quer nos bons quer nos maus momentos, principalmente durante os últimos tempos.

Obrigada.

Resumo

O ribossoma eucariótico é maior e mais complexo do que o bacteriano, sendo constituído por rRNA e por 80 proteínas ribossomais, as quais são essenciais para a montagem, estrutura e função do ribossoma. A maioria das proteínas são extremamente conservadas a nível de estrutura e função ao longo da evolução, evidenciando a sua importância. Neste grupo está incluída a proteína ribossomal S4 cujo gene é encontrado em autossomas em todos os vertebrados exceto nos mamíferos, os quais possuem uma cópia ligada ao cromossoma X - *RPS4X*. Em algumas linhagens está presente também uma cópia ancestral no cromossoma Y - *RPS4Y*. Nos primatas, a duplicação do gene *RPS4Y* originou uma segunda cópia, a *RPS4Y2*, nos Macacos do Velho Mundo. No murganho (*Mus musculus*), a *Rps4x* adquiriu um parálogo no cromossoma 6, a *Rps4l*, sem intrões e que codifica uma proteína muito semelhante à original. Ambos os genes, *RPS4Y2* e *Rps4l*, são expressos durante a espermatogénese e poderão compensar o silenciamento transcricional da cópia ligada ao X durante a inativação meiótica dos cromossomas sexuais (IMCS).

O presente estudo visa a identificação e caraterização de retrogenes autossómicos potencialmente funcionais da RPS4 em diferentes linhagens de mamíferos. Para tal combinámos análises filogenéticas, estruturais e funcionais *in silico*, de forma a averiguar se estes retrogenes mantêm a função da RPS4X e a poderão compensar.

Identificámos potenciais ortólogos da S4l apenas em roedores murídeos, cricetídeos e numa espécie da família *Spalacidae*, o que indica que o evento da duplicação que originou este retrogene ocorreu há mais de 43,9 milhões de anos. A taxa de evolução dos retrogenes S4l identificados em roedores, calculada através da razão dN/dS, indica uma baixa acumulação de substituições não sinónimas e portanto conservação da sequência ancestral depois da duplicação. Não foram detetadas alterações estruturais que possam comprometer a função da proteína codificada por estes retrogenes, reforçando a hipótese de que poderão compensar a S4x nesta linhagem que perdeu a cópia ligada ao cromossoma Y.

Nas restantes espécies de mamíferos analisadas que retêm genes *RPS4* ligados ao Y (gato e humano) não foram encontrados duplicados noutros cromossomas, tal como no

porco, em que ainda não se sabe se existe o gene *RPS4Y*. No coelho, no cão, na vaca e no cavalo foram encontradas sequências com potencial funcional. A taxa de evolução (dN/dS) para estes retrogenes denota que as sequências tendem a conservar o seu estado ancestral, à exceção do retrogene do cromossoma 9 da vaca que parece não estar sujeito a pressões seletivas (dN/dS≈1).

Quanto à análise estrutural, apesar das alterações na sequência da proteína codificada pelos retrogenes relativamente à ancestral do X causarem algumas modificações nas interações aminoacídicas intra-moleculares, não são detetáveis alterações estruturais globais, dentro dos limites que a resolução da estrutura modelo disponível permite. No caso do retrogene do cromossoma 6 do cavalo, detetou-se uma *β-sheet* mais curta e o impacto desta alteração é desconhecido.

Em suma, na maioria das espécies analisadas em que não existe uma cópia da S4 no cromossoma Y identificamos retrogenes autossómicos da *RPS4X* potencialmente funcionais (murganho, rato e outros roedores, cão, vaca e coelho), candidatos a compensar a função da RPS4X durante a IMCS na espermatogénese.

Palavras-chave: Cromossomas sexuais; Proteínas ribossomais; Duplicação de genes; Inativação meiótica; Espermatogénese.

Abstract

The eukaryotic ribosome is bigger and more complex than the bacterial one. It is constituted by rRNA and 80 ribosomal proteins which are essential for ribosome assembly, structure and function. Most proteins are highly conserved in terms of structure and function throughout evolution, which highlights their importance. Ribosomal protein S4 (RPS4) is included in this group, encoded by a gene found on the autosomes in all vertebrates except mammals, which have an X-linked copy – *RPS4X*. In some lineages an ancestral copy on the Y chromosome – *RPS4Y* – *is* also present. In primates, a duplication of the *RPS4Y* gene originated a second copy, *RPS4Y2*, in Old World Monkeys. In the mouse (*Mus musculus*), *Rps4x* acquired an intronless paralogue on chromosome 6, *Rps4l*, encoding a very similar protein to the original. Both genes, *RPS4Y2* and *Rps4l*, are expressed during spermatogenesis and may compensate for the transcriptional silencing of the X-linked copy during meiotic sex chromosome inactivation (MSCI).

The present study aims at the identification and characterization of potentially functional autosomal retrogenes of RPS4 in different mammalian lineages. To this purpose we combined phylogenetic, structural and functional *in silico* analysis, in order to investigate whether these retrogenes have kept RPS4X function and may eventually compensate for its silencing.

S4L potential orthologues were identified only in murine rodents, *Cricetidae* and in a species of the *Spalacidae*, dating the duplication event that gave rise to this retrogene at more than 43.9 million years.

The rate of evolution of the S4l retrogenes identified in rodents, calculated through dN/dS ratios, indicates a low accumulation of non-synonymous substitutions and, therefore, the conservation of ancestral sequences after duplication. Structural changes that may compromise the function of the encoded protein by these retrogenes were not detected, reinforcing the hypothesis that they may compensate the S4x of this lineage, where the Y-linked copy was lost.

In other mammalian species that were analyzed and which retain the Y-linked *RPS4* genes (cat and human), no autosomal duplicates were found, as well as in pig, a species where *RPS4Y* has not been identified yet. In rabbit, dog, cow and horse, potentially functional paralogues were found. The rate of evolution (dN/dS) for these retrogenes indicates that the sequences tend to conserve their ancestral state, except for the retrogene found on chromosome 9 of the cow, which does not appear to be under selective pressure (dN/dS≈1).

Despite of the changes in the sequence of the proteins encoded by these retrogenes, which cause some changes in the amino acid intramolecular interactions comparatively to the ancestral X-linked protein, no global changes to the global tridimensional structure of the proteins are detectable, within the limits of the resolution that the model available allow. The retrogene on chromosome 6 of horse has a shorter *β-sheet* than the X-linked gene and the impact of this change is unknown.

In summary, we identified *RPS4X* potentially functional autosomal retrogenes (mouse, rat and other rodents, dog, cow and rabbit) in most of the analyzed species that do not retain the S4 copy on Y chromosome, which makes them good candidates to compensate for RPS4X function during IMCS in spermatogenesis.

Key-words: Sex chromosomes; Ribosomal proteins; Gene duplication; Meiotic sex chromosome inactivation; Spermatogenesis.

Índice

Lista de tabelas

Lista de figuras

Lista de abreviaturas

Bp	Pares de bases
cDNA	DNA complementar
CDS	Sequência codificante
DNA	Ácido desoxirribonucleico
ICX	Inativação do cromossoma X
IMCS	Inativação meiótica dos cromossomas sexuais
mRNA	RNA mensageiro
ORF	Grelha de leitura aberta
RNA	Ácido ribonucleico
RPS4	Proteína ribossomal S4
RPS4X	Proteína ribossomal S4 ligada ao X
RPS4Y	Proteína ribossomal S4 ligada ao Y
rRNA	Ácido ribonucleico ribossómico

Aminoácido	Abreviatura	Letra
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Ácido aspártico	Asp	D
Ácido Glutâmico	Glu	E
Cisteína	Cys	C
Fenilalanina	Phe	F
Glicina	Gly	G
Glutamina	Gln	Q
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Prolina	Pro	P
Serina	Ser	S
Tirosina	Tyr	Y
Treonina	Thr	T
Triptofano	Trp	W
Valina	Val	V

Introdução

1. Cromossomas sexuais em mamíferos

A reprodução sexuada sendo amplamente comum entre a maior parte dos seres vivos, apresenta mecanismos de determinismo sexual muito diversos, entre eles o sistema genético XY. Este sistema altamente diferenciado está presente em grande parte dos mamíferos placentários (*Eutéria*), sendo os machos caracterizados por heterogamia (cariótipo XY) e as fêmeas por homogamia (cariótipo XX)[1-4].

Os cromossomas sexuais, apesar da sua origem comum, a partir de um par de autossomas homomórficos ancestrais, são distinguíveis morfológica e geneticamente, uma vez que evoluíram independentemente[5-7]. Como resultado desse processo diferencial, um dos autossomas adquiriu um gene - *SRY* e posteriormente diferenciou-se no cromossoma Y[8]. Este gene codifica um fator de transcrição envolvido no desenvolvimento sexual masculino, através do controlo da atividade de outros genes. Assim, a presença de um cromossoma Y determina o sexo masculino[9,10].

A acumulação de alelos vantajosos para o sexo masculino com funções específicas na espermatogénese e determinismo sexual[2,10-12] na proximidade do *SRY* terá originado uma supressão progressiva da recombinação entre os cromossomas X e Y. Nos mamíferos placentários atuais a recombinação entre os dois cromossomas sexuais, durante a meiose masculina, resume-se a duas pequenas regiões, as quais contêm *loci* homólogos. Estas regiões situam-se na extremidade do braço curto e do braço longo destes cromossomas e são denominadas regiões pseudoautossómicas ou *PARs*[13,14]. A ausência de recombinação levou a que ao longo da evolução a região não recombinante no cromossoma Y acumulasse mutações e deleções que resultaram na perda de genes ancestrais ligados ao cromossoma Y, sendo uma grande parte deste cromossoma heterocromática[15]. Contrariamente ao cromossoma Y, o cromossoma X exibe um tamanho bastante maior e um conteúdo genético muito diverso.[2].

Como resultado do fenómeno anteriormente descrito, no sexo heterogamético (XY), existem genes ligados ao cromossoma X que perderam o seu homólogo funcional no cromossoma Y. Assim, haveria uma diferença na dosagem entre machos e fêmeas, a qual é atenuada por mecanismos de compensação de dose[16] que nos mamíferos, envolvem a inativação aleatória de um dos cromossomas X (ICX) em cada célula, nas fêmeas[17]. Esta inativação é conseguida pelo aumento da síntese de um RNA não-codificante denominado XIST, o qual se acumula no cromossoma que será inativado, revestindo-o e levando à aquisição de marcas epigenéticas de repressão, que incluem a modificação das histonas e a metilação do DNA, culminando no silenciamento

genético do cromossoma[18]. Desta forma garante-se a igualdade da quantidade de produtos genéticos de *loci* ligados aos cromossomas sexuais em ambos os sexos, sendo o estado inativo em cada célula herdado de forma estável, gerando indivíduos adultos com células que expressam o X paterno e outras o X materno.[6,16,19] A existência de dois cromossomas heterólogos no sexo masculino leva ainda a que um outro fenómeno ocorra durante a espermatogénese: a inativação meiótica dos cromossomas sexuais (IMCS). Através deste processo, há o silenciamento ao nível da transcrição de genes ligados a ambos os cromossomas (X e Y) [20-22] na fase de paquíteno na meiose, depois do emparelhamento dos cromossomas homólogos autossómicos maternos e paternos. Pensa-se que esta inativação meiótica se deve à ausência de emparelhamento dos cromossomas sexuais[23], prevenindo-se assim o início de eventos de recombinação entre regiões não homólogas do X e do Y, os quais poderiam causar aneuploidia e infertilidade nas gerações seguintes [21,22].

2. Proteínas ribossomais ligadas aos cromossomas sexuais: o caso da S4X, S4Y1 e S4Y2

O ribossoma eucariota é constituído por 2 subunidades (60S e 40S), sendo cada uma composta por moléculas de rRNA e uma grande quantidade de proteínas ribossomais[24,25]. A maioria das proteínas ribossomais são extremamente conservadas durante a evolução, sugerindo que estas desempenham funções importantes. Assim, apesar de o rRNA ser uma molécula catalítica, as proteínas ribossomais são cruciais para uma tradução eficiente, pois ajudam na estabilização de estruturas específicas de rRNA em subunidades ribossomais maduras e na montagem do ribossoma, promovendo a correta dobragem de rRNAs[25-27].

Neste estudo vamos focar-nos na proteína S4, localizada na interface das subunidades 40S/60S da subunidade pequena do ribossoma[27] sendo extremamente conservada, tanto a nível de sequência como função e posição[28]. O gene da *RPS4* encontra-se em autossomas na maioria dos vertebrados, como a galinha, em que se situa no cromossoma 4[29], mas ligado ao X em mamíferos (*RPS4X*)[30]. Dado o elevado grau de homologia entre a RPS4 da galinha e a RPS4X de mamíferos, pensa-se que esta desempenha um papel importante no desenvolvimento em vertebrados[29]. Em algumas linhagens de mamíferos placentários e em marsupiais[31] o gene *RPS4* tem uma cópia

ancestral ligada ao Y (*RPS4Y1*) expressa em todos os tecidos[29,32]. Existem evidências de que este gene se encontra no bloco *X-degenerate*[7] (versão degenerada do autossoma ancestral que deu origem ao cromossoma Y[11]), sugerindo assim que a *RPS4Y1* estava presente antes da radiação dos Eutéria[33]. No entanto, noutros mamíferos, como por exemplo nos clados *Rodentia*, *Carnivora*, *Artiodactyla* (vaca) *e ainda na espécie Equus caballus*, o gene *RPS4Y1* foi perdido devido à degeneração do Y durante a evolução[30,33,34]. No clado *Carnivora*, a espécie *Felis catus* representa uma exceção porque contém duas cópias da S4 ligadas ao cromossoma Y[35]. Em primatas foi descoberta uma segunda cópia ligada ao Y, à qual foi atribuído o nome *RPS4Y2*[7], uma vez que foi originada a partir da duplicação do gene ancestral *RPS4Y1*, depois da divergência dos Macacos do Novo Mundo, mas antes da radiação dos Macacos do Velho Mundo[30].

Os genes *RPS4X* e *RPS4Y* codificam isoformas diferentes de proteínas ribossomais S4 mas estas são permutáveis e pensa-se que ambas são essenciais para o desenvolvimento, dado que uma expressão insuficiente da S4 poderá estar associada com a síndrome de Turner[32,36]. A cópia *RPS4Y2* é especificamente expressa em próstata e nas células da linha germinativa em testículo [30,37] na linhagem humana, o que sugere que esta desempenha um papel importante na espermatogénese.

Nos roedores, mais especificamente em murganho (*Mus musculus*), a *Rps4x* adquiriu um parálogo autossómico, localizado no cromossoma 6, sem intrões e expresso predominantemente em testículo – a *Rps4l*[38]. Visto que o gene *Rps4l* é desprovido de intrões, assume-se que tenha sido originado por retrotransposição do gene *RPS4* ligado ao cromossoma X durante a evolução, tal como foi enunciado para outras proteínas ribossomais com cópias autossómicas[39]. A retrotransposição é um processo através do qual o mRNA é reversamente transcrito em cDNA e depois inserido novamente no genoma, numa posição diferente, originando uma retrocópia[40]. No rato (*Rattus norvegicus*), também existe um homólogo da S4 nos autossomas, cujo produto terá sido detetado por análise de proteómica nos ribossomas de testículo[38]. A observação destes padrões de expressão em cópias autossómicas levanta a hipótese que estas possam ter um papel importante na compensação dos genes parálogos ligados ao X, os quais estão inativos durante a espermatogénese, devido à condensação dos cromossomas sexuais na meiose masculina[38,39].

3. Evolução de genes duplicados

No genoma da maior parte dos organismos, a estreita semelhança entre genes sugere que os mesmos apresentam uma história evolutiva comum, estando também relacionados em termos estruturais e funcionais (exemplo das proteínas ribossómicas S4X, Y1 e Y2). Muitas vezes, estes genes formam famílias de genes parálogos, constituindo um dos exemplos de como novos genes podem surgir a partir da duplicação de genes ancestrais[41].

A duplicação génica é um mecanismo através do qual um gene dá origem a dois genes parálogos, exatamente iguais. Para que ambos sejam mantidos no genoma por seleção, têm de divergir entre si de alguma forma[42]. Então, uma das cópias fica livre de constrangimentos funcionais e, assim, está sujeita a seleção, deriva e pode ainda acumular mutações que possivelmente a levarão a desenvolver uma nova função, enquanto a outra cópia continua a desempenhar a função original[43-45].

Segundo Ohno, apesar de causarem alterações de dosagem e expressão de genes entre espécies, as duplicações génicas são essenciais para a diversificação evolutiva dos genomas dos organismos e podem ser produzidas por *crossing over* desigual, retrotransposição, transposição de DNA duplicado e poliploidização[40,42,46].

As diferentes cópias podem seguir trajetos distintos, tais como: (i) silenciamento - ocorre frequentemente devido à acumulação de mutações deletérias as quais levam à degeneração e perda funcional do duplicado, transformando-o num pseudogene; (ii) neofuncionalização - um dos duplicados adquire uma nova função a qual é benéfica para o organismo e portanto perdura pela ação da seleção natural, e ainda, (iii) subfuncionalização - ambos os duplicados se complementam, dividindo as funções entre si sendo a sua função conjunta a mesma que a do gene ancestral[41,46,47]. A subfuncionalização aumenta o tempo de retenção dos duplicados no genoma podendo estes evoluir posteriormente para realizar novas funções[48].

Os genes duplicados que analisamos neste estudo surgiram por retrotransposição. Este processo origina duplicados sem intrões nem elementos reguladores (como o caso dos promotores), sem os quais a retrocópia não tem qualquer viabilidade funcional. Então, tende a acumular mutações que a levam a transformar-se num pseudogene, sendo também designado de pseudogene processado, dado que se formou por retrotransposição de um RNA maduro[49]. Se uma retrocópia for inserida perto de um promotor já existente, ou recrutar novos elementos reguladores de outros genes a montante, pode tornar-se funcional[40,50]. Os retrogenes podem ser introduzidos em

regiões favoráveis à sua transcrição, como por exemplo, próximo de outras regiões codificantes ou em intrões de sequências codificantes tendo, por isso, maior probabilidade de serem expressos do que os retrogenes que são introduzidos longe desse tipo de sequências[50-52]. Como a retrotransposição ocorre na linha germinal, as retrocópias são frequentemente introduzidas em genes expressos durante esse processo ou muito perto dos mesmos, já que a cromatina está aberta nessa região. Isto pode justificar que muitos dos retrogenes autossómicos, derivados do cromossoma X, apresentem padrões de expressão específicos de testículo[51,53,54].

No nosso estudo centrámo-nos em retrogenes de proteínas ribossomais, as quais por serem conservadas podem ser utilizadas para identificação de espécies. A presença de retrogenes com elevada identidade genética com o gene pode levar a erros de análise, dado poder ocorrer co-hibridização dos primers em ambas as sequências. Daí a importância de caraterizar estas sequências, que podem em muitos casos ser específicas de determinadas linhagens.

Para quantificar as pressões seletivas que atuam sobre regiões codificadoras de proteínas frequentemente é usada a razão dN/dS, a qual compara a taxa de substituições sinónimas e não sinónimas entre sequências[55].

Objetivos

A questão que queremos abordar neste trabalho é a seguinte:
Os retrogenes autossómicos da RPS4X poderão compensar a sua função durante a inativação meiótica dos cromossomas sexuais na espermatogénese?

Se assim for esperamos encontrar retrogenes funcionais em diferentes linhagens de mamíferos.

Assim, com este estudo pretende-se identificar retrogenes autossómicos da *RPS4* em mamíferos, os quais apresentem potencial funcional de forma a poderem codificar proteínas que sejam compensatórias dos seus homólogos ligados ao X quando estes últimos se encontram silenciados durante a IMCS na espermatogénese. Além disso, pretende-se situar o evento de duplicação que originou a S4l em roedores e seguir a sua história evolutiva em mamíferos.

Objetivos específicos

1. Determinar a idade aproximada da duplicação que originou o retrogene *Rps4l* identificado em murganho e analisar a sua história evolutiva;

2. Identificar retrogenes autossómicos da RPS4X potencialmente funcionais, em diferentes linhagens de mamíferos;

3. Determinar o potencial funcional dos retrogenes encontrados através de análises evolutivas, funcionais e estruturais *in silico.*

Materiais & Métodos

1. Pesquisa e caraterização de retrogenes

Para identificar ortólogos da *Rps4l* de murganho noutras espécies fez-se um *Nucleotide Blast* (disponível em http://blast.ncbi.nlm.nih.gov/Blast.cgi) com a sequência de cDNA da *Rps4l* de *M. musculus* (anotada na **UniProt** com o código Q3V1Z5, disponível em http://www.uniprot.org/uniprot/Q3V1Z5) na base de dados *Reference genomic sequences (refseq_genomic)* do NCBI, escolhendo no parâmetro *Optimize for* "*Highly similar sequences (megablast)*".

Para a pesquisa de retrogenes autossómicos da *RPS4X* em mamíferos, recorreu-se ao **Ensembl** (http://www.ensembl.org/index.html) usando como sequência de pesquisa o CDS da *Rps4x* de *M. musculus,* a qual está bem anotada e estudada (sequência modelo em todo o estudo). No parâmetro "*Orthologues*" selecionaram-se os mamíferos placentários e destes elegeram-se apenas as espécies que apresentam o genoma sequenciado, definido em cromossomas e cujo gene que codifica a proteína RPS4X se encontra efetivamente ligado a esse cromossoma. Nos casos do rato, do coelho e do cão, estavam descritas sequências denominadas *RPS4X* mas noutros cromossomas pelo que recorremos ao **NCBI** (http://www.ncbi.nlm.nih.gov/) para procurar as sequências da *RPS4* ligadas ao cromossoma X. Deu-se ainda prioridade na análise àquelas espécies para as quais fosse possível obter tecidos para estudos laboratoriais posteriores e humano (*R. norvegicus*, *O. Cuniculus*, *C. l. familiaris*, *F. catus*, *E. caballus*, *B. taurus*, *S. scrofa* e *H. sapiens).*

Para cada espécie, fez-se um Blat da *RPS4X* contra o respetivo genoma, recorrendo ao **UCSC Genome Browser** (http://genome.ucsc.edu/cgi-bin/hgBlat). O único gene parálogo autossómico conhecido em *M. musculus*, a *Rps4l*, tem uma identidade nucleotídica de 79,8% com o gene *Rps4x*. Assim, selecionaram-se todas as sequências até 750 bp e com uma percentagem de identidade até 75% para que fossem identificados apenas retrogenes (sem sequência intrónica) não demasiado divergentes da *RPS4X* e com potencial para manter a sequência aminoacídica. Depois de selecionadas estas sequências, mantiveram-se para análise posterior apenas aquelas que mantinham os 4 domínios proteicos íntegros, descritos na base de dados **InterPro** para a sequência modelo (disponível em http://www.ebi.ac.uk/interpro/protein/P62702 com o código de acesso P62702). As sequências que apresentavam um codão STOP depois dos 4 domínios proteicos foram também incluídas no estudo. Para aquelas em que por Blat não se recuperou a sequência completa da região codificante do gene

(CDS) procurou-se a sequência em falta além daquela que foi identificada, sendo que em alguns casos a sequência total se tornou mais longa relativamente à da *RPS4X* respetiva.

Para as sequências recolhidas do Blat começadas em "ATG", fez-se o seu alinhamento com a respetiva *RPS4X* de forma a verificar se coincidia com o codão de iniciação. Algumas destas sequências quando traduzidas *in silico* apresentavam codões de terminação prematuros e foram portanto excluídas das análises posteriores. As sequências que apresentavam ORF foram mantidas para as análises seguintes.

Nos casos em que por Blat não se obteve a sequência completa da região codificante do gene (CDS), foi necessário fazer uma pesquisa manual do "ATG" a 5' ou do codão STOP a 3' para além da sequência que foi identificada sendo, por vezes, necessário fazer o alinhamento com a sequência total do gene *RPS4X* para verificar se os nucleótidos encontrados corresponderiam a parte de um exão ou a uma região intrónica que pudesse ter sido incorporada no retrogene. Nos casos em que faltava o "ATG", começou-se por procurar o primeiro codão de iniciação na sequência resultante do Blat, fazendo depois o alinhamento com o CDS da *RPS4X* da espécie respetiva, de forma a verificar se havia perda de sequência significativa. Caso ocorresse perda de sequência (frequentemente eram perdidos os exões iniciais), procurou-se um "ATG" a montante do início da sequência que tinha sido identificada por Blat. Seguiu-se então, o alinhamento da nova sequência (com o acréscimo de bp) com o CDS da *RPS4X*, para verificar se o "ATG" encontrado seria um codão de iniciação e se a sequência apresentava um quadro de leitura aberto - destas sequências apenas as que apresentavam ORF prosseguiram para análise.

Nos casos em que não estava presente um codão de terminação procedeu-se da mesma forma na direção 3'.

A sintenia é a conservação da ordem dos genes entre segmentos de cromossomas de duas ou mais espécies. Recorrendo aos mapas de sintenia disponíveis no ***Ensembl*** procurou-se identificar sequências localizadas na região sinténica à da *Rps4l* do murganho.

2. Análises filogenéticas e evolutivas

As análises evolutivas foram conduzidas usando a **versão 6 do Mega**[56]. As sequências foram alinhadas usando o algoritmo *Muscle* e seguidamente construiu-se uma árvore filogenética de máxima verosimilhança. Para tal, escolheu-se no parâmetro *Models* a opção *"Find Best DNA/Protein Models"* para encontrar o melhor modelo aplicável às sequências em análise.

Para situar o evento da duplicação que originou o retrogene *Rps4l* em roedores, foi utilizada a base de dados **TimeTree** (http://www.timetree.org/).

Calculou-se a relação dN/dS usando o programa **PAL2-NAL** (http://www.bork.embl.de/pal2nal/#RunP2N), de forma a quantificar as pressões evolutivas que atuam sobre as proteínas (ortólogos da *Rps4l* e retrogenes dos mamíferos).

3. Anotação funcional dos retrogenes encontrados: conservação de domínios e estrutura prevista das proteínas

Para determinar se os ortólogos da *Rps4l* encontrados serão potencialmente funcionais, fez-se o alinhamento das sequências traduzidas recorrendo ao algoritmo *Muscle* do ***Geneious Pro v.5.5.8***[57], sendo que, como referência, foi utilizada a RPS4X de *M. musculus*, visto que é uma espécie extremamente bem estudada, ao contrário dos roedores encontrados que ainda não têm o genoma sequenciado e/ou agrupado em cromossomas. As substituições aminoacídicas encontradas entre os retrogenes e a RPS4X foram analisadas recorrendo ao **SIFT** (http://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html), o qual se baseia no grau de conservação dos resíduos aminoacídicos, alinhando a sequência em análise com sequências estreitamente relacionadas, recolhidas por PSI-BLAST. De forma a confirmar a informação sobre o impacto das substituições aminoacídicas na estabilidade e função dos duplicados encontrados (ortólogos da RPS4L e retrogenes em mamíferos) recorreu-se ainda ao **PROVEAN Protein** (http://provean.jcvi.org/seq_submit.php), o qual faz a

previsão do impacto de uma substituição aminoacídica ou de uma deleção na função biológica de uma proteína utilizando um alinhamento mais extenso relativamente ao **SIFT** e tendo em conta também as propriedades químicas dos aminoácidos envolvidos.

A conservação de determinadas posições na proteína ao longo da evolução, indica a importância das mesmas na sua função. Então, de forma a identificar quais as posições da RPS4 mais conservadas ao longo da evolução das espécies, procedeu-se ao alinhamento de várias sequências da mesma, representativas de diferentes grupos filogenéticos. As alterações previstas como não toleradas em posições conservadas da RPS4 podem causar modificações da estrutura terciária da proteína e daí advir consequências a nível funcional.

Para completar esta análise modelou-se a estrutura terciária dos ortólogos encontrados bem como da RPS4L recorrendo ao **Swiss-model** (http://swissmodel.expasy.org/)**.** Como molde, este programa utilizou o modelo de RPS4 de *O. cuniculus*[58], o único modelo de S4 de mamífero disponível na base de dados **Protein Data Bank** (**PDB)** com o código de acesso 4KZX (disponível em http://www.rcsb.org/pdb/explore/explore.do?structureId=4kzx). Este modelo tem no entanto uma resolução relativamente baixa (7.8 Å) e não permite determinar com precisão as interações entre átomos. Por último, recorrendo ao **PyMOL**[59] (http://pymol.org/academic), compararam-se as interações polares entre aminoácidos dentro de cada molécula, nos ortólogos e na RPS4X já anotada (de *M. musculus*), para as posições conservadas na RPS4 com substituições não toleradas, de forma a avaliar o potencial para estas sequências terem uma função semelhante ao respetivo parálogo no cromossoma X.

Após o alinhamento das sequências dos duplicados identificados pelos métodos anteriores e da respetiva RPS4X para os mamíferos selecionados, usando o ***Geneious Pro v.5.5.8***[57]*,* procedeu-se da mesma forma (descrita para os ortólogos da RPS4L) para averiguar o potencial funcional dos retrogenes encontrados. Com a exceção de que no **Swiss-model** (http://swissmodel.expasy.org/) foi necessário modelar a RPS4X de mamífero (sequência conservada em todos os mamíferos analisados), para fazer a comparação entre as interações polares dos seus aminoácidos e as interações polares entre os aminoácidos nos duplicados autossómicos.

Resultados

1. Pesquisa e caraterização de órtologos da RPS4L

1.1. Pesquisa de ortólogos RPS4L

A *Rps4l* de murganho é o único retrogene autossómico da *Rps4x* que está descrito e relativamente bem caraterizado (**Figura 1**) sendo expresso de forma específica em testículo, durante a espermatogénese.[38]

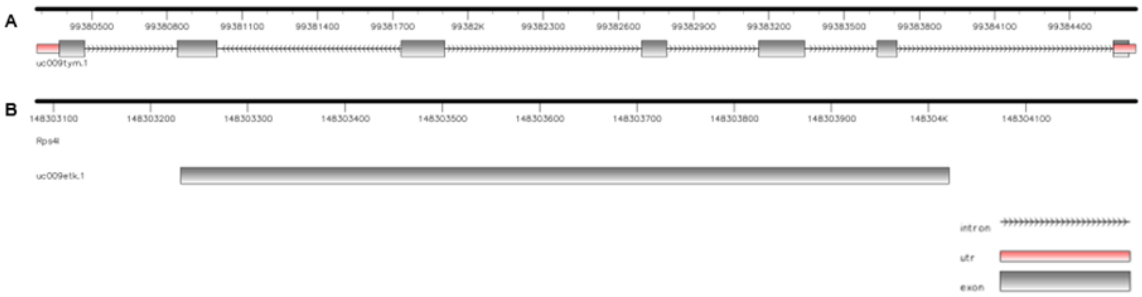


Figura 1 - Estrutura da *Rps4x* (A) e retrogene *Rps4l* (B) de *M. musculus*. Na sequência genómica da *Rps4x* encontram-se anotados os exões e os intrões bem como as UTR's (A). As sequências pertencem ao *Genome browser, assembly July 2007 (NCBI/mm9),* sendo a representação esquemática feita através do software fancyGENE (http://bio.ieo.eu/fancygene/).

A identificação dos ortólogos da *Rps4l* de murganho noutras espécies feita através de um *Nucleotide Blast* resultou nas sequências da **Tabela 1**. Através da análise da árvore de distância genética dos resultados do Blast, disponível na opção *"Distance tree of results"* (**Figura A do Material suplementar**), é evidente que das sequências identificadas apenas as de roedores são potencialmente ortólogas da *Rps4l* de murganho. Deste resultado foram eleitas apenas as sequências da classe *Rodentia*, com uma cobertura de sequência de 100% e identidade superior a 80%.

Tabela 1 - Sequências da RPS4L analisadas neste estudo.

Espécie	Sequência nucleotídica	Sequência aminoacídica	Segmentos genómicos
M. musculus	NR_003634.2	-	NW_001030820.1
***R. norvegicus* (RPS4Y2)**	NM_001109612.1	NP_001103082.1	NW_007905798.1
P. m. bairdii	XM_006986665.1	XP_006986727.1	NW_006501337.1
C. griseus	XM_003505484.1	XP_003505532.1	NW_003614152.1
M. auratus	XM_005072944.2	XP_005073001.1	NW_004801640.1
M. ochrogaster	XM_005364620.1	XP_005364677.1	NW_004949097.1
N. galili	XM_008837614.1	XP_008835836.1	NW_008341884.1

Códigos de acesso das sequências nucleotídica (coluna 1) e proteica (coluna 2) da RPS4L, encontradas por Blast. NM e NP sequências manualmente curadas; XM e XP sequências modelo previstas pela análise da sequência genómica; NW segmentos genómicos (*scaffolds*); NR refere-se a RNA.

Devido ao seu interesse em estudos laboratoriais, o murganho (*Mus musculus*) e o rato (*Rattus norvegicus*) são organismos modelo cujo genoma está sequenciado e bem anotado, permitindo a pesquisa *in silico* de regiões sinténicas entre os cromossomas de ambas. Então, foi possível determinar que o retrogene de rato identificado na pesquisa por Blast é de facto ortólogo da *Rps4l* de murganho, através da análise posicional e de conservação da sintenia. No **Ensembl**, está descrita a sintenia entre a região 148,354,656-148,355,598 do cromosoma 6 de murganho e a região 182,745,448-182,746,381 do cromossa 4 de rato - **Figura 2.** Comparando as duas regiões, observa-se a localização da *Rps4l* num dos intrões do gene *Tmtc1* nas duas espécies, confirmando-se, então, que este é de facto o ortólogo da *Rps4l* em *R. norvegicus* - **Figura 3**.

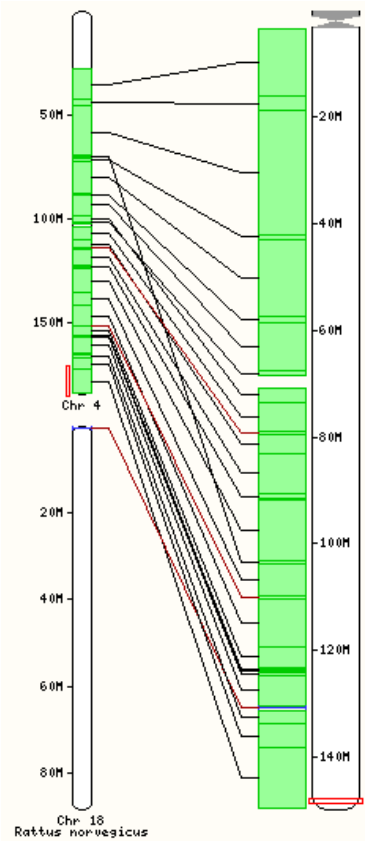


Figura 2 - Mapa de sintenia entre *Mus musculus* e *Rattus norvegicus*. O cromossoma 6 de *Mus musculus* (direita) partilha regiões sinténicas no cromossoma 18 e 4 de *Rattus norvegicus* (esquerda). A barra vermelha em cada um dos cromossomas representa a posição da *Rps4l*, designada *Rps4y2* em *R. norvegicus*. Esquema disponível em: http://www.ensembl.org/Mus_musculus/Location/Synteny?db=core&r=6%3A148354656-148355598&g=ENSMUSG00000063171&t=ENSMUST00000071745&otherspecies=Rattus_norvegicus.

Figura 3 - Comparação das regiões sinténicas de murganho e rato que contêm a *Rps4l*. (A) Cromossoma 6 de *Mus musculus* e (B) Cromossoma 4 de *Rattus norvegicus*. O retrogene *Rps4l* encontra-se num dos intrões do gene *Tmtc1* em ambas as espécies.

Dado que nas espécies de roedores com possíveis ortólogos da *Rps4l* os respetivos genomas não estão ainda bem anotados e os genes/segmentos genómicos (*contigs*) não estão agrupados em cromossomas, recorreu-se apenas à análise posicional, ou seja, verificou-se sempre que as sequências encontradas se localizavam num intrão do gene *Tmtc1*. Na **Tabela 2** estão descritas as percentagens de identidade nucleotídica e similaridade aminoacídica com a RPS4L e RPS4X de murganho, das sequências de roedores selecionadas para análise.

Tabela 2 - Comparação das sequências nucleotídicas e aminoacídicas da RPS4 de roedores.

Espécie	*Rps4l* (IN)	RPS4l (SA)	*Rps4x* (IN)	RPS4X (SA)
R. norvegicus	96,5%	99,6%	79,9%	92,8%
P. m. bairdii	93,5%	98,9%	79,8%	93,6%
C. griseus	91,3%	99,2%	77,7%	93,6%
M. auratus	91%	99,2%	77,9%	92,8%
M. ochrogaster	90,9%	99,2%	78,2%	93,2%
N. galili	81,6%	85,6%	70,2%	80,4%

Valores determinados por alinhamento com o CDS/proteína de *M. musculus* (RPS4L e RPS4X - sequência de referência); todos os potenciais retrogenes da *Rps4l* apresentam 100% de cobertura e 789 bp de comprimento. IN - identidade nucleotídica; SA - similaridade aminoacídica.

Tanto a nível nucleotídico como aminoacídico, a sequência mais semelhante à *Rps4l* de *M. musculus* é a de *R. norvegicus* com 96,5% e 99,6%. Todas apresentam identidades acima dos 90%, indicando conservação da sequência, exceto a de *N. galili* com 81,6% e 85,6%, respetivamente (**Tabela 2**).

1.2. Análise filogenética e evolutiva da RPS4L

Dado que não foram encontradas sequências homólogas noutros mamíferos além de roedores, é provável que a S4L seja específica desta classe. Com o intuito de investigar há quanto tempo terá ocorrido a duplicação que a originou e para delinear a sua história evolutiva, construíram-se duas árvores filogenéticas com as sequências selecionadas: uma com sequências nucleotídicas e outra com sequências aminoacídicas - **Figura 4 e Figura 5**, respetivamente. Apesar da falta de conhecimento sobre o genoma dos roedores analisados, encontraram-se sequências anotadas como S4X, as quais foram incluídas nas árvores filogenéticas. Também as RPS4 de outras espécies mais divergentes das estudadas foram adicionadas a cada árvore (grupos externos ou *outgroups*): *G. gallus*, *D. melanogaster* e *D. rerio*, o que permitiu identificar a raiz da árvore filogenética.

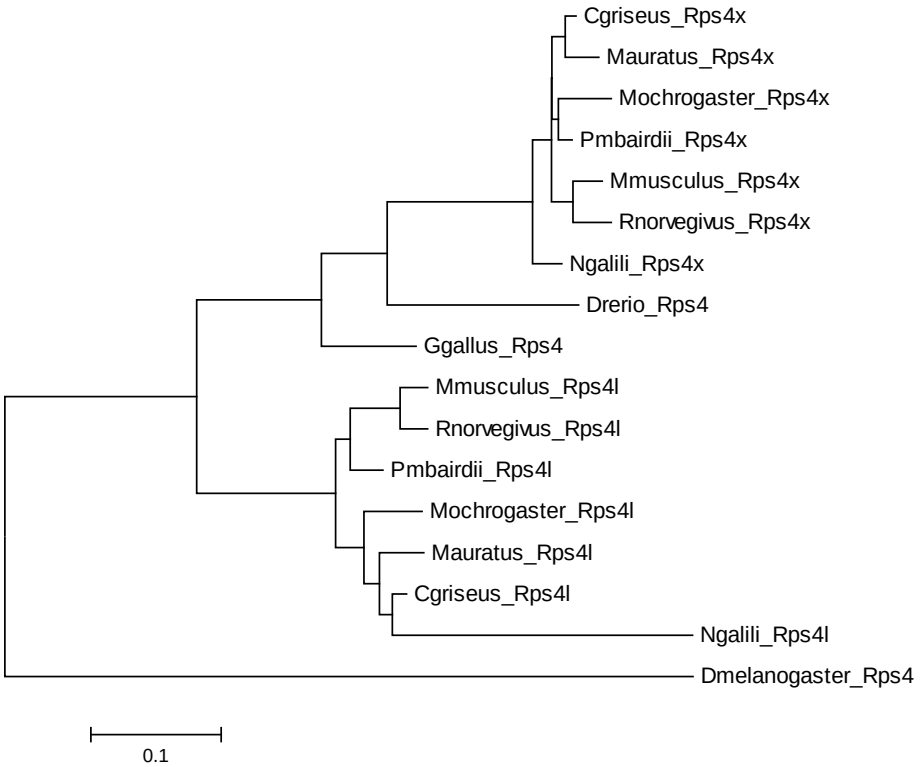


Figura 4 - Árvore filogenética das sequências nucleotídicas dos ortólogos da *Rps4l* de *M. musculus e respetivas Rps4x. A árvore foi* obtida através do Mega6 com os parâmetros T92+G (Tamura 3-parameter + Gamma distribution). O *outgroup* mais divergente é a *Drosophila melanogaster*. Códigos de acesso para *Rps4x*: *M. musculus:* NM_009094.1; *R. norvegicus*: NM_001007600.2; *C. griseus*: NM_001246673.1; *M. auratus*: XM_005081210.1; *M. ochrogaster*: XM_005360537.2; *P. m. bairdii*: XM_006996049.1; *N. galili*: XM_008855556.1.

Pelo agrupamento das *Rps4x* verificado na árvore filogenética (**Figura 4**), denota-se a semelhança entre todas elas e com as *Rps4x* de murganho e rato, as quais estando melhor estudadas fornecem suporte à sua classificação como *Rps4* efetivamente ligada ao cromossoma X nos organismos estudados. Também as sequências potencialmente ortólogas da *Rps4l* identificadas por Blast agrupam juntas, sendo a de *N. galili* a mais distante de todas as outras.

A relação filogenética entre as sequências, tanto da *Rps4x* como da *Rps4l* reflete o padrão de evolução das espécies (**Figura B do Material suplementar**), exceto para as sequências de *G. gallus* e *D. rerio*. O facto de na árvore da **Figura 4** a sequência do peixe zebra (*D. rerio*) e não da galinha (*Aves*) estar filogeneticamente mais próxima dos mamíferos poderá ser explicada pelo facto da S4 da galinha ser mais divergente. Segundo Zinn et al.[29], a RPS4 da galinha tem quatro aminoácidos diferentes relativamente à S4X humana, três dos quais estão também presentes na cópia da S4 ligada ao cromossoma Y em humanos, o que leva a crer que o gene S4 da galinha possa ter surgido por recombinação entre as duas sequências – S4X e S4Y – o que pode explicar a sua divergência relativamente às RPS4X das outras espécies.

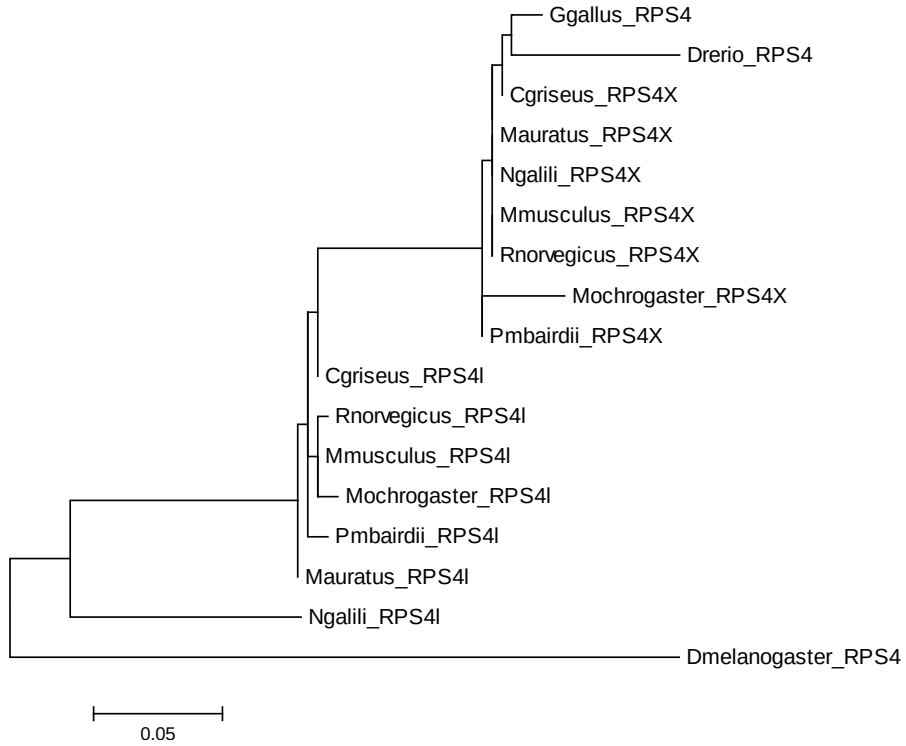


Figura 5 - Árvore filogenética das sequências aminoacídicas dos ortólogos da RPS4L de *M. musculus e respetivas* RPS4X, obtida através do Mega6 com os parâmetros LG+G (LG model + Gamma distribution). O *outgroup* mais divergente é a *Drosophila melanogaster.* Códigos de acesso para RPS4X: *M. musculus*: NP_033120.1; *R. norvegicus*: NP_001007601.1; *C. griseus*: NP_001233602.1; *M. auratus*: XP_005081267.1; *M. ochrogaster*: XP_005360594.1; *P. m. bairdii*: XP_006996111.1; *N. galili*: XP_008853778.1.

Na árvore construída com as sequências aminoacídicas (**Figura 5**) continua a verificar-se a existência de dois grupos de sequências: o grupo das RPS4X e o grupo das RPS4L. No grupo das RPS4X, as sequências de *M. ochrogaster* e de *P. m. bairdii* são as mais divergentes saindo num ramo diferente de todas as outras RPS4X que são originadas de um ramo comum, do qual diverge, posteriormente, a sequência de *C. griseus*. Quanto ao grupo das RPS4L, a sequência de *N. galili* continua a ser a mais divergente seguida da sequência de *M. auratus*. Destacam-se as sequências de *R. norvegicus*, *M. musculus* e *M. ochrogaster* originadas a partir do mesmo ramo. Novamente se reflete o padrão de evolução das espécies confirmado pela **Figura B do Material suplementar**, excetuando-se as espécies *G. gallus* e *D. rerio,* pelo enunciado anteriormente aquando da análise da **Figura 4**.

Segundo a base de dados **TimeTree**, o ancestral comum mais recente entre as espécies mais divergentes de roedores que apresentam *Rps4l* (Taxon A: *M. musculus* e Taxon B: *N. galili*), terá existido há pelo menos 43,9 milhões de anos. Assim a duplicação que originou este retrogene deverá ter ocorrido antes.

De modo a avaliar as pressões evolutivas que atuam sobre os retrogenes S4L identificados em roedores, calculou-se a razão entre as substituições sinónimas e não sinónimas (dN/dS) entre cada retrogene e o gene ancestral do cromossoma X da espécie respetiva (**Tabela 3**).

Tabela 3 - Taxas de substituições sinónimas e não sinónimas entre a RPS4X e respetivos duplicados em roedores.

Espécie	N	S	dN	dS	dN/dS
M. musculus	586,5	199,5	0,0362	3,2115	0,0113
R. norvegicus	577,6	208,4	0,0387	2,8238	0,0137
P. m. bairdii	594,1	191,9	0,0322	3,6557	0,0088
C. griseus	608,5	177,5	0,0347	23,5746	0,0015
M. auratus	600,3	185,7	0,0317	19,9006	0,0016
M. ochrogaster	614,1	171,9	0,0484	18,8431	0,0026
N. galili	613,1	172,9	0,1081	58,7279	0,0018

Resultados obtidos com o PAL2NAL. N - número de sítios não sinónimos; S - número de sítios sinónimos; dN - taxa de substituições não sinónimas; dS - taxa de substituições sinónimas.

A razão dN/dS presente na **Tabela 3** apresenta valores muito baixos, inferiores à unidade para todas as espécies, indicando um maior número de substituições sinónimas do que não sinónimas desde que ocorreu a duplicação. Ou seja, as pressões evolutivas que têm atuado sobre as sequências visam a conservação do seu estado ancestral - seleção negativa.

Calculámos ainda a mesma razão entre pares de ortólogos S4L para as espécies de roedores. Dividimos as espécies em dois grupos filogeneticamente mais próximos (**Figura 4**) para minimizar a ocorrência de substituições múltiplas na mesma posição. Testámos as espécies que agrupam no primeiro ramo (*M. musculus*, *R. norvegicus* e *P. m. bairdii*) e no segundo ramo (*M. ochrogaster, M. auratus, C. griseus* e *N. galili*) – **Tabela 4**.

Tabela 4 - Taxas de substituições sinónimas e não sinónimas entre os ortólogos da RPS4L das espécies de roedores.

Espécie	N	S	dN	dS	dN/dS
M. musculus vs R. norvegicus	662,1	123,9	0,0015	0,2577	0,0059
M. musculus vs P. m. bairdii	693,4	92,6	0,0059	0,7427	0,0079
R. norvegicus vs P. m. bairdii	689,2	96,8	0,0074	0,7211	0,0103
M. ochrogaster vs M. auratus	705,5	80,5	0,0057	0,7848	0,0073
M. ochrogaster vs C. griseus	715,9	70,1	0,0028	0,8740	0,0032
M. ochrogaster vs N. galili	706,5	79,5	0,0701	56,3191	0,0012
M. auratus vs C. griseus	701,7	84,3	0,0029	0,4944	0,0058
M. auratus vs N. galili	709,5	76,5	0,0702	2,3769	0,0295
C. griseus vs N. galili	716,2	69,8	0,0697	2,0691	0,0337

Resultados obtidos com o PAL2NAL. N - número de sítios não sinónimos; S - número de sítios sinónimos; dN - taxa de substituições não sinónimas; dS - taxa de substituições sinónimas.

Para os pares de ortólogos da RPS4L também se verifica uma razão dN/dS muito baixa, inferior à unidade em todos os casos (**Tabela 4**), ou seja, as substituições sinónimas prevalecem relativamente às não sinónimas mesmo entre os duplicados.

1.3. Análise funcional e estrutural *in silico* da RPS4L

1.3.1. Substituições não toleradas e anotação de domínios

Para avaliar o potencial da RPS4L de murganho e dos seus ortólogos noutras espécies para manter conservada a função do gene parálogo ligado ao cromossoma X durante a inativação meiótica dos cromossomas sexuais, fez-se a comparação dos retrogenes encontrados com a RPS4X de referência (murganho). Para este efeito alinhou-se a RPS4X com as RPS4L das diferentes espécies de roedores (**Figura 6**), de modo a identificar as posições com substituições e determinar se estas estão conservadas na RPS4 ao longo da evolução (**Figura 7**). As posições que são mantidas intactas ao longo da evolução das espécies serão de extrema importância a nível da função da proteína no ribossoma.

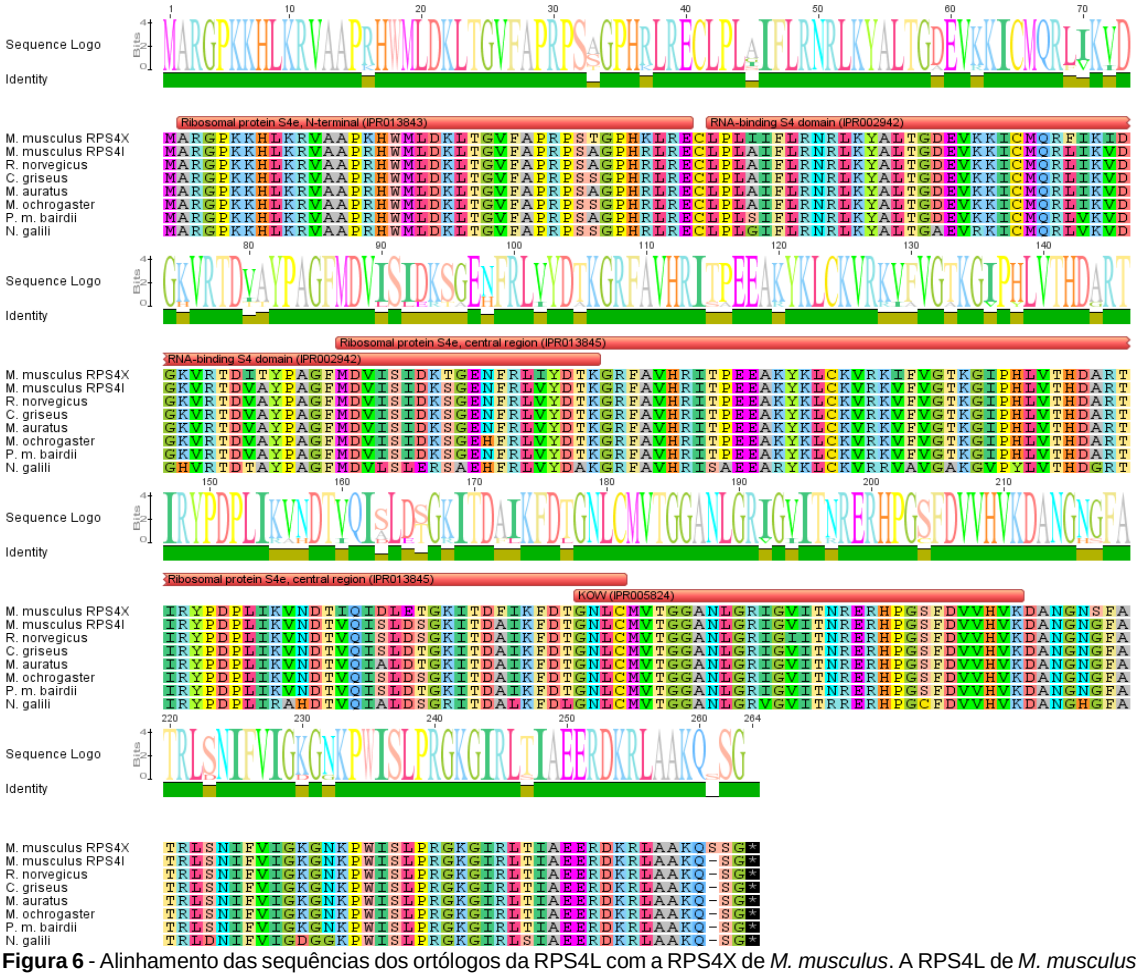


Figura 6 - Alinhamento das sequências dos ortólogos da RPS4L com a RPS4X de *M. musculus*. A RPS4L de *M. musculus* tem 17 alterações aminoacídicas relativamente à RPS4X da mesma espécie, apresentando um aminoácido a menos (263). A vermelho estão representados os domínios anotados na base de dados InterPro para a RPS4X - *N-terminal:* 3-39, *RNA-binding:* 42-106, *Central region:* 87-181 e *KOW:* 178-211.

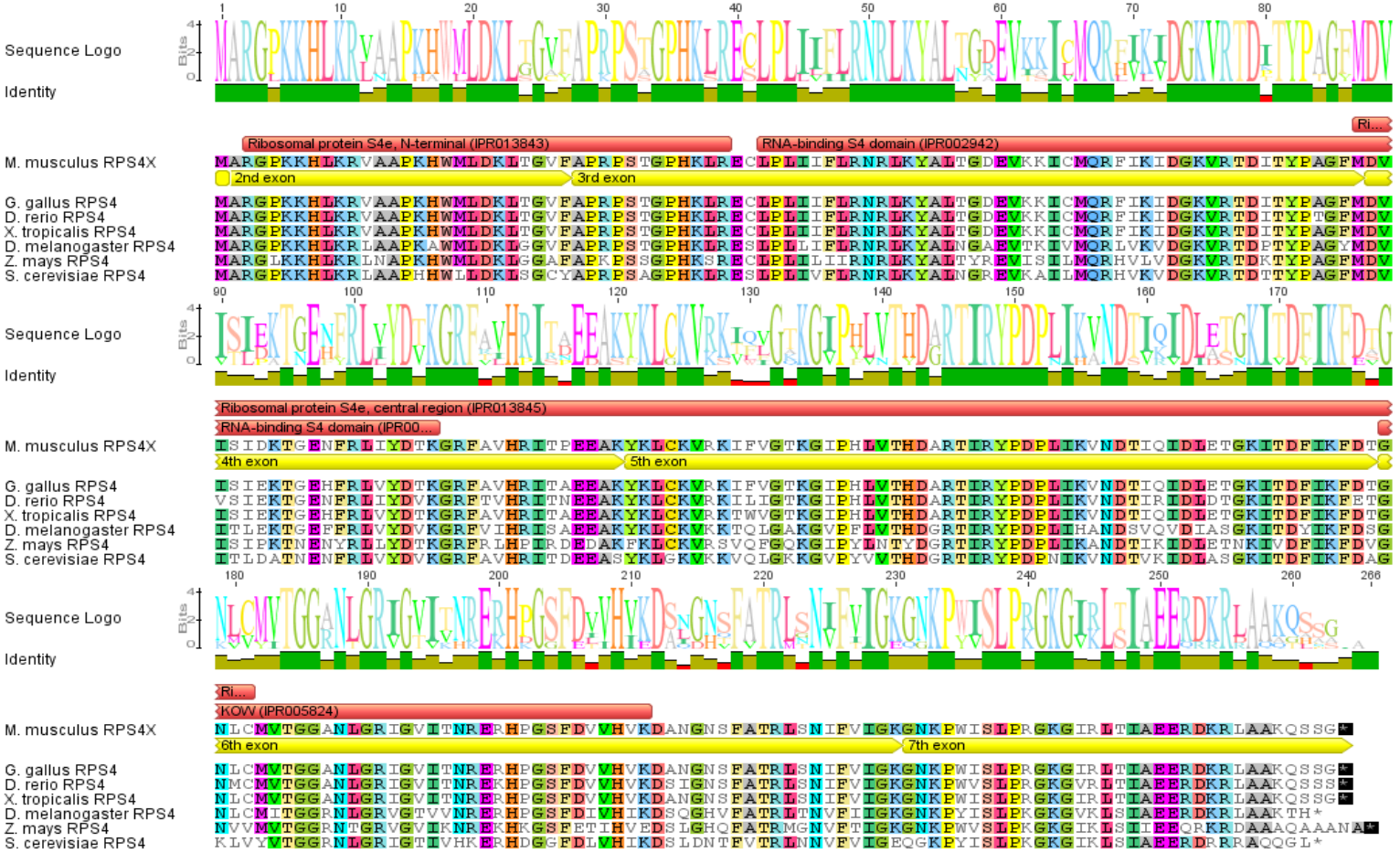


Figura 7 - Alinhamento das RPS4 de organismos representando vários clados, de forma a avaliar a conservação da proteína ao longo da evolução. Pela observação do alinhamento, verifica-se que a proteína ribossómica S4 é evolutivamente conservada desde os mamíferos até às leveduras, passando pelas aves, peixes, anfíbios, insetos e plantas. A amarelo estão representados os exões da RPS4X. Cada espécie do alinhamento representa um clado (entre parêntesis estão representados os códigos de acesso das sequências): *M. musculus* – mamíferos; *G. gallus* (NP_990439.1) – aves; *D. rerio* (NP_001005589.1) – peixes; *X. tropicalis* (NP_988912.1) – anfíbios; *D. melanogaster* (NP_001287055.1) – insetos; *Z. mays* (NP_001266498.1) – plantas; *S. cerevisiae* (NP_012073.1) – fungos.

Através do alinhamento das RPS4 de diferentes clados (**Figura 7**) determinou-se as substituições aminoacídicas em cada domínio funcional da proteína (**Tabela 5**) de forma a averiguar o grau de conservação dos mesmos**.**

Tabela 5 - Conservação dos domínios da RPS4 ao longo da evolução.

Domínio	Nº de substituições aminoacídicas/Nº de aminoácidos do domínio	% de domínio com substituições
N-terminal (IPR013843) Posição 3 - 40	4/37	10,81%
RNA-binding (IPR002942) Posição 42 - 106	17/65	26,15%
Central region (IPR013845) Posição 87 - 181	27/95	28,42%
KOW (IPR005824) Posição178 - 211	10/34	29,41%

O domínio mais conservado é o *N-terminal*, apresentando menor percentagem de alteração e o menos conservado é o *KOW* com 10 substituições.

Como as posições menos conservadas deverão ter menor importância a nível funcional, para a análise estrutural apenas foram consideradas as alterações em posições conservadas. Assim, apesar de ainda não se saber qual a sua função pensa-se que, dada a sua elevada conservação, o primeiro domínio (*N-terminal*) desempenhe um papel importante a nível funcional na proteína, provavelmente na interação com o ribossoma. Quanto ao segundo domínio, conforme a anotação presente na **Gene Ontology** com o acesso GO: 0003723 (disponível em http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0003723), está envolvido na interação seletiva e de forma não covalente com uma molécula de RNA ou parte dela. Os últimos domínios exibem uma percentagem de alteração mais elevada relativamente aos outros dois e não apresentam, ainda, função descrita.

De seguida fez-se uma previsão *in silico* do impacto das substituições encontradas na função da proteína com o algoritmo **SIFT**, classificando-as em toleradas ou não-toleradas (**Tabela 6**), o que foi confirmado recorrendo ao **PROVEAN Protein** (**Tabela 7**).

Tabela 6 - Substituições não toleradas na RPS4L de diferentes espécies de roedores, em relação à RPS4X de *M. musculus*.

Substituições não toleradas na RPS4L de roedores					
Domínio	Posição	Conservação na RPS4[(a)]	RPS4X	RPS4L	Aminoácidos tolerados*
RNA-binding S4 (IPR002942)	45	NC	I	A	m f S G L V I
	75	**C**	**K**	**H**	**R K**
	90	NC	I	L	V I
	95	**C**	**T**	**S**	**T**
	96	NC	G	A	s K D N G
S4e, central region (IPR013845)	163	**C**	**D**	**A**	**g E N S D**
	168	**C**	**K**	**R**	**Q K**
	172	NC	F	A	S Y F
	173	**C**	**I**	**L**	**A V I**
KOW (IPR005824)	204	NC	S	C	y i v p l d H t n q G e A R K S

[(a)]Posições determinadas no alinhamento da Figura 6; Posições não toleradas por espécie: *R. norvegicus*, *C. griseus* e *M. ochrogaster* - 3 (45, 95, 172); M. auratus - 4 (45, 95,163, 172); *P. m. bairdii* - 2 (95, 172); N. galili - 9 (75, 90, 95, 96, 163, 168, 172, 173, 204). *As letras maiúsculas indicam aminoácidos representados em alguma das espécies no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão. A bold estão as substituições em posições conservadas da RPS4.

Das 52 posições com substituições aminoacídicas observadas no alinhamento dos duplicados dos vários roedores (**Figura 6**), apenas 10 correspondem a substituições não toleradas e dessas, somente 5 pertencem a posições conservadas ao longo da evolução das espécies (**Figura 7**). Apesar de a comparação ter sido feita relativamente à RPS4X de *M. musculus*, confirmou-se que as posições analisadas contêm o mesmo aminoácido nas RPS4X da espécie respetiva (**Figura C do Material suplementar**).

Tabela 7 - Resultados do PROVEAN Protein para as posições não toleradas detetadas pelo SIFT nos ortólogos da RPS4L em roedores.

Espécie	Posição	Provean Protein
R. norvegicus	95	Deletéria
C. griseus		
M. ochrogaster		
P. m. bairdii		
M. auratus	95	Deletéria
	163	Deletéria
N. galili	75	Deletéria
	95	Deletéria
	163	Deletéria
	168	Neutra
	173	Neutra

Os programas são coerentes na maior parte dos resultados, ou seja, as posições não toleradas pelo **SIFT** são também identificadas como deletérias pelo **PROVEAN Protein**, uma vez que os dois se baseiam no alinhamento de sequências que apresentam elevada semelhança com a sequência em questão, para prever o impacto das substituições aminoacídicas. No entanto, existem duas exceções, em *Nannospalax galili*, em que, para as posições 168 e 173, o **SIFT** classifica a substituição como não tolerada mas o **PROVEAN Protein**, enuncia que a mesma substituição é neutra. Tal facto pode ocorrer devido ao primeiro algoritmo fazer uma previsão da tolerância com base num PSI-BLAST, o qual usa apenas sequências proteicas ao passo que o segundo algoritmo faz um alinhamento mais extenso, com um maior número de sequências traduzidas e portanto, é mais provável que encontre o aminoácido substituído entre as sequências alinhadas. Além disso, o **PROVEAN Protein** tem em conta também as propriedades químicas dos aminoácidos em questão ao fazer a análise da substituição ocorrida.

O ortólogo de *N. galili* é o mais divergente, apresentando nove alterações não toleradas, três das quais também deletérias relativamente à RPS4X de referência (**Tabela 6** e **Tabela 7**), sendo cinco em posições conservadas da RPS4 (**Figura 7**), o que poderá resultar numa alteração da função da proteína. Também comparativamente à RPS4X, em *P. m. bairdii* existem duas alterações aminoacídicas não toleradas e em *R. norvegicus, C. griseus e M. ochrogaster* existem três, sendo que, em todas elas, apenas uma das alterações ocorreu numa posição conservada do alinhamento das RPS4 (**Figura 7**) – posição 95, que é também considerada deletéria (**PROVEAN Protein**) – **Tabela 6** e **Tabela 7**. Quanto à espécie *M. auratus*, situa-se entre as sequências menos divergentes em relação à RPS4X, possuindo quatro alterações aminoacídicas não toleradas, sendo duas em posições conservadas da RPS4 (**Figura 7**) e consideradas deletérias (**PROVEAN Protein**) – **Tabela 6** e **Tabela 7**.

As substituições não-toleradas e deletérias (**SIFT** e **PROVEAN Protein**), em posições conservadas ao longo da evolução das espécies (**Figura 7**) foram analisadas estruturalmente, devido à possibilidade de causarem modificações na estrutura terciária da proteína e concomitantemente na sua função.

1.3.2. Análise estrutural das RPS4L por modelação comparativa

O cerne desta análise consiste em verificar as alterações nas interações aminoacídicas para as posições conservadas que apresentam substituições aminoacídicas não toleradas entre a sequência original que sabemos ser funcional (RPS4X) e os seus duplicados. Para tal, construíram-se modelos da estrutura da RPS4X de *M. musculus* e da RPS4L das diferentes espécies de roedores, por modelação comparativa, utilizando o modelo da RPS4 de coelho determinado experimentalmente já disponível no **PDB** e comparou-se as interações entre aminoácidos, para as posições que diferem entre os duplicados e a RPS4X.

Das 17 alterações aminoacídicas verificadas aquando do alinhamento da RPS4X com a RPS4L em *M. musculus* (**Figura 6 e Tabela I do Material suplementar**) e das 18 verificadas aquando do alinhamento da RPS4X com a RPS4L em *R. norvegicus* (**Figura D e Tabela II do Material suplementar**), apenas 3 correspondem a substituições não toleradas e dessas apenas uma corresponde a uma posição conservada da RPS4 (Figura 7). Em murídeos (*M. musculus*, *R. norvegicus*) – **Tabela 8** – e em quase todos os cricetídeos (*P. m. bairdii, C. griseus* e *M.ochrogaster*) a RPS4L apresenta apenas uma alteração não tolerada e deletéria (**Tabela 6** e **Tabela 7**) em relação à RPS4X, numa posição conservada (95), em que uma treonina é substituída por uma serina.

Tabela 8 - Substituições não toleradas entre a RPS4X e o seu duplicado, a RPS4L, em murídeos (*M. musculus* e *R. norvegicus*).

Substituições não toleradas na RPS4L de murídeos					
Domínios	Posição	C/NC relativamente à RPS4	RPS4X	Chr6/Chr4.1[1] (RPS4L)	Aminoácidos tolerados*
RNA-binding S4 (IPR002942)	45	NC	I	A	m f S G L V I
	95	**C**	**T**	**S**	**T**
S4e, central region (IPR013845)	172	NC	F	A	S Y F

Nota: As RPS4L do murganho e do rato foram analisadas em conjunto dado que as duas sequências são idênticas. [1]Localização cromossómica em *M. musculus*/Localização cromossómica em *R. norvegicus*. *As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão.

As interações previstas para esta posição são mantidas com os mesmos aminoácidos comparativamente com a RPS4X (isoleucina-92, ácido aspártico-93, glicina-96 e ácido glutâmico-97) – **Figura 8A e 8B** – e portanto, não deverão haver repercussões a nível estrutural.

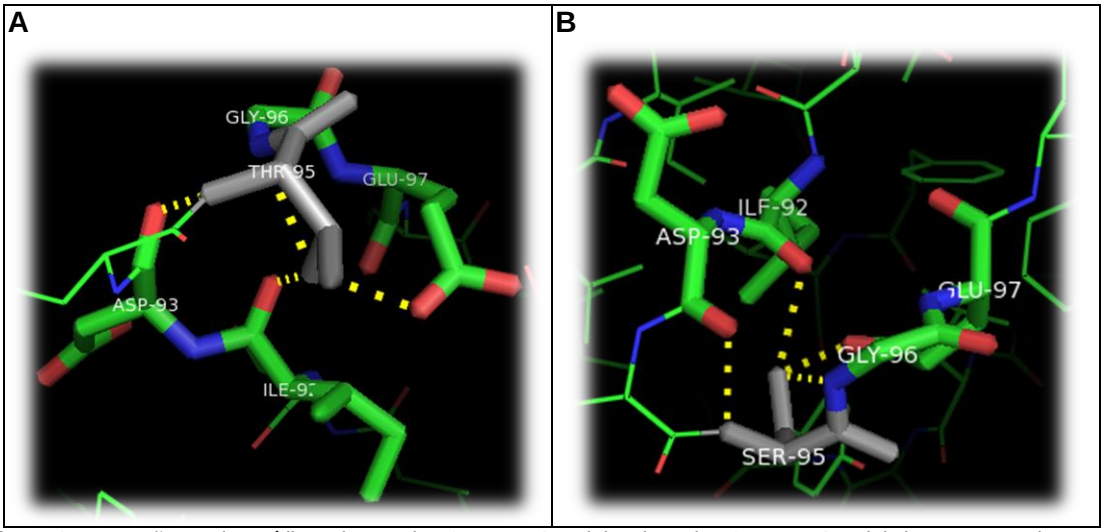


Figura 8 - Interações aminoacídicas da treonina-95 na RPS4X (A) e da serina-95 na RPS4L (B) de *Mus musculus*. Para determinar se houve alterações nas interações aminoacídicas na RPS4L relativamente à RPS4X em murídeos (*M. musculus*, *R. norvegicus*) e em cricetídeos (*P. m. bairdii, C. griseus* e *M.ochrogaster*) foram comparadas as interações polares na única posição conservada em que se verificou uma substituição não tolerada nestas espécies, utilizando o modelo da proteína de murganho. A cinza está representado o aminoácido alterado entre as duas sequências e o tracejado amarelo representa as interações polares entre aminoácidos. Como se pode verificar as interações nesta posição não sofrem alterações.

A RPS4L de ***M. auratus*** apresenta duas substituições aminoacídicas em posições conservadas (**Tabela 6**), uma sendo comum entre o segundo e terceiro domínio e a segunda específica do terceiro. Para a posição 95 não se observam alterações nas interações aminoacídicas relativamente à RPS4X, como descrito para as espécies anteriores (**Figura 8B**). Quanto à posição 163, há a substituição de um ácido aspártico na RPS4X, o qual estabelece interações polares com a treonina-166, glicina-167 e lisina-168 (**Figura 9A**) por uma alanina, no ortólogo, que estabelece interações apenas com a glicina-167 e a lisina-168 (**Figura 9B**).

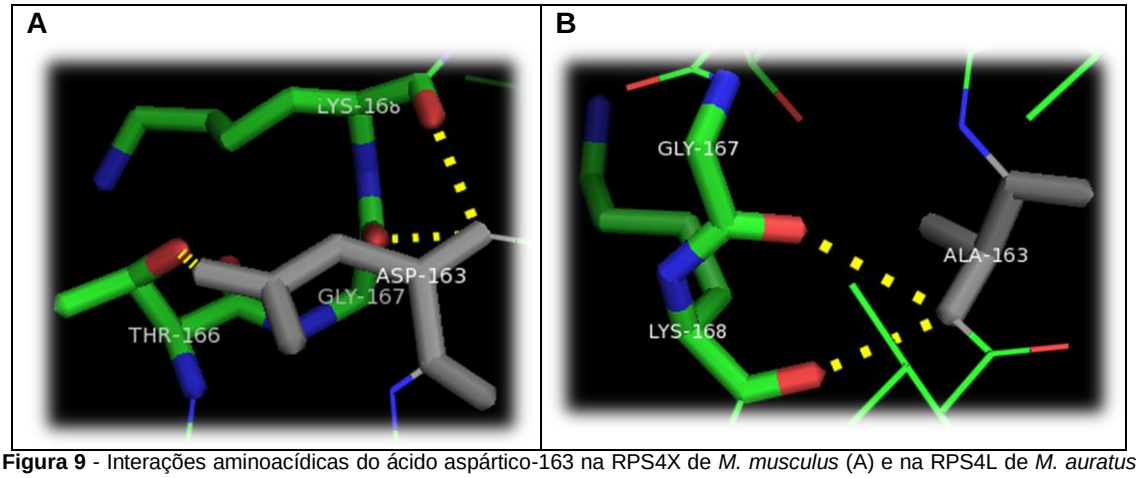


Figura 9 - Interações aminoacídicas do ácido aspártico-163 na RPS4X de *M. musculus* (A) e na RPS4L de *M. auratus* (B). Na RPS4X de *M. musculus* o ácido aspártico-163 estabelece três interações polares (amarelo tracejado) enquanto na RPS4L de *M. auratus* a alanina-163 estabelece apenas duas. A cinza está representado o aminoácido alterado.

De forma a avaliar se a alteração descrita na **Figura 9** para a posição 163 poderá causar modificações a nível estrutural, fez-se a sobreposição das estruturas terciárias da RPS4X de *M. musculus* e da RPS4L de *M. auratus*. Observando a **Figura 10** e tendo em conta a baixa resolução possível, não se verificam alterações estruturais significativas, dado que apenas uma das ligações é afetada, portanto, pensa-se que esta modificação não terá influência a nível funcional.

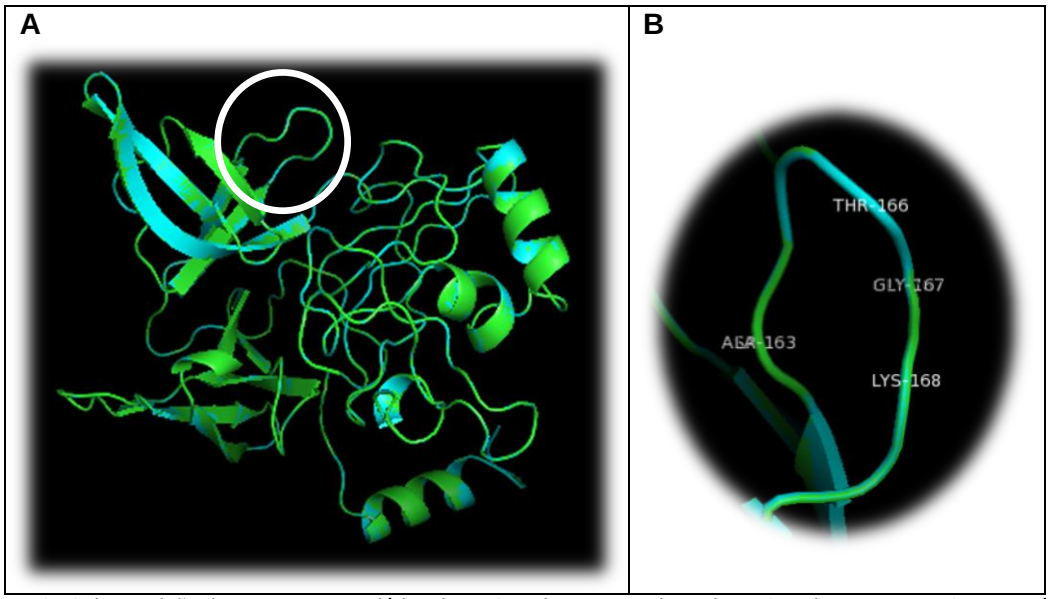


Figura 10 - Sobreposição das estruturas terciárias da RPS4X de *M. musculus* e da RPS4L de *M. auratus*. Com um círculo branco (Figura 10A) encontra-se assinalada a posição que difere entre as duas estruturas bem como as respetivas interações, o que pode ser visto a pormenor na Figura 10B. A verde está retratada a RPS4L e a azul a RPS4X.

A RPS4L de ***N. galili*** (*Spalacidae*) apresenta várias alterações não toleradas em posições conservadas, e muitas sobrepõem-se no mesmo domínio funcional. Assim é

muito provável que haja uma alteração da estrutura da proteína e torna-se mais improvável que esta mantenha a função, portanto não foi considerada para análise estrutural.

2. Pesquisa e anotação de retrogenes autossómicos da *RPS4X*

2.1. Pesquisa de retrogenes da S4X em diferentes linhagens de mamíferos

A pesquisa de ortólogos da *RPS4X* de *M. musculus* em mamíferos no **Ensembl** e no **NCBI** resultou nas sequências da **Tabela 9**. O gato, apesar de ainda não ter o genoma agrupado em cromossomas, foi incluído neste estudo porque se sabe que possui 2 cópias no Y, tal como o Homem[7,32,35].

Tabela 9 - Sequências da RPS4 ligadas aos cromossomas sexuais nas espécies analisadas neste estudo.

Espécie	Sequência nucleotídica	Sequência proteica
M. musculus	NM_009094.1	NP_033120.1
R. norvegicus	NM_001007600.2	NP_001007601.1
O. cuniculus	XM_002720107.2	XP_002720153.1
C. l. familiaris	NM_001252042.1	NP_001238971.1
F. catus	XM_004000627.2	XP_004000676.1
B. taurus	NM_001035445.2	NP_001030522.1
S. scrofa	NM_001204283.1	NP_001191212.1
E. caballus	NM_001163950.1	NP_001157422.1
H. sapiens		
RPS4X	NM_001007.4	NP_000998.1
RPS4Y1	NM_001008.3	NP_000999.1
RPS4Y2	NM_001039567.2	NP_001034656.1

Códigos de acesso das sequências nucleotídica (coluna 1) e proteica (coluna 2) da S4X para os mamíferos estudados e, no caso da espécie humana, das cópias presentes no cromossoma Y (códigos NCBI). NM e NP – sequências manualmente curadas; XM e XP – sequências modelo previstas pela análise da sequência genómica.

Como resultado do Blat com o CDS da *RPS4X* de cada organismo no genoma da própria espécie, obteve-se uma lista de várias sequências, as quais foram manualmente filtradas de forma a identificar apenas retrogenes (sem sequência intrónica) potencialmente funcionais. As espécies estudadas e o número de sequências analisadas para cada espécie estão indicadas na **Figura 11**.

Figura 11 - Número de sequências estudadas por espécie. Número total de sequências resultantes do Blat com o CDS da *RPS4X* da respetiva espécie no seu próprio genoma (1ª coluna); número de sequências que foram analisadas seguidamente como potenciais retrogenes (2ª coluna); e ainda, na 3ª coluna, o número de sequências que cumprem os restantes parâmetros estabelecidos (percentagem de identidade com a *RPS4X* e presença de ORF), as quais foram posteriormente analisadas.

A seleção foi feita segundo os seguintes parâmetros:

(i) Tamanho: Como a *RPS4X* apresenta um CDS de 792 bp, selecionaram-se todas as sequências com um mínimo de 750 bp e até 850 bp, para que não fossem excluídos, da primeira pesquisa, retrogenes com inclusão de pequenas sequências intrónicas ou com pequenas deleções (**Tabela III do Material suplementar**);

(ii) Percentagem de identidade: foram selecionadas sequências com identidade até 75% com a *RPS4X*, para que fossem incluídos não só os retrogenes mais recentes ou com sequência mais conservada mas também outros que, não sendo demasiado divergentes, pudessem ainda assim apresentar diferenças na sequência aminoacídica (**Tabela III do Material suplementar**);

(iii) ORF (Grelha de leitura aberta ou *open reading frame*) - Das sequências anteriores mantiveram-se apenas aquelas que não apresentavam codões STOP prematuros (**Tabela 10**) dentro dos 4 domínios proteicos descritos para a RPS4X de *M. musculus* na base de dados **InterPro** (disponível em http://www.ebi.ac.uk/interpro/protein/P62702 com o código P62702).

Tabela 10 - Retrogenes da RPS4X com grelha de leitura aberta (ORF).

Espécie	Cópias autossómicas	% de identidade nucleotídica	% de similaridade aminoacídica	Comprimento da sequência (aminoácidos)
Mus musculus	Chr 7	97,0%	94,3%	264
	Chr 6 (RPS4l)	80,3%	93,2%	263
	Chr 7.1	78,8%	83,4%	256
Rattus norvegicus	Chr 4	100%	100%	264
	Chr 4.1	79,5%	92,8%	263
Oryctolagus cuniculus	Chr 1	94,9%	98,5%	264
	Chr 12	99,3%	98,1%	264
	Chr14	98,2%	97,0%	264
	Chr12.1*	92,8%/93,1%	84,8%	269
Canis lupus familiaris	Chr8*	95,6%/95,6%	98,9%	268
	Chr3	96,8%	93,2%	221
	Chr29	96,8%	92,8%	221
Bos taurus	Chr23	94,2%	99,2%	264
	Chr 9	99,0%	98,9%	263
	Chr 7	98,9%	98,5%	264
	Chr 8	97,0%	96,2%	264
	Chr 10	96,9%	95,1%	264
	Chr3*	97,8%/97,5%	85,2%	250
Equus caballus	Chr6	91,3%	92,5%	264
	Chr18*	85,3%/71,2%	75,9%	222

Estão representados os duplicados resultantes da seleção descrita anteriormente (critérios i, ii, iii), a sua localização cromossómica, bem como a respetiva percentagem de identidade nucleotídica e similaridade aminoacídica com a S4X da mesma espécie e ainda o tamanho da cadeia polipeptídica. Estes retrogenes apresentam potencial para codificar uma proteína funcional. *Sequências em que a região codificante obtida por Blat não estava completa e em que foi necessário procurar a restante sequência - nestas indica-se a % de identidade do Blat/ % após recuperação manual da sequência.

Para as sequências recolhidas do Blat (**Tabela 10**) começadas em “ATG”, fez-se o seu alinhamento com a respetiva *RPS4X* de forma a verificar se coincidia com o codão de iniciação. Algumas destas sequências quando traduzidas *in silico* apresentavam codões de terminação prematuros e foram portanto excluídas das análises posteriores (**Figura 12**). Nos casos em que por Blat não se obteve a sequência completa da região codificante do gene (CDS), foi necessário fazer uma pesquisa manual do “ATG” a 5’ ou do codão STOP a 3’ da sequência, como descrito nos métodos. Apresenta-se um exemplo de uma sequência na qual estava em falta o “ATG”, em que se começou por procurar o primeiro codão de iniciação na sequência do Blat, fazendo depois o

alinhamento com o CDS da *RPS4X* da espécie respetiva, de forma a verificar se havia perda de sequência significativa - **Figura 13A**. Como faltava a sequência correspondente aos exões iniciais da *RPS4X*, procurou-se um "ATG" a montante do início da sequência que tinha sido identificada por Blat sendo que, no exemplo apresentado, apenas foi necessário procurar o número de bases em falta aquando da observação do alinhamento da **Figura 13B**. Seguiu-se então, o alinhamento da nova sequência (com o acréscimo de bases) com o CDS da *RPS4X*, para verificar se o "ATG" encontrado seria um codão de iniciação e posteriormente, verificou-se se apresentava uma grelha de leitura aberta - **Figura 13B**. Nos casos em que não estava presente um codão de terminação (exemplo na **Figura 14A**) procedeu-se da mesma forma na direção 3' - **Figura 14B**.

Após anotação manual, apenas as sequências que apresentavam ORF prosseguiram para análise. As que apresentavam nucleótidos em falta a 3' ou possuíam um codão STOP prematuro ficaram com um tamanho diferente da sequência obtida por Blat, resultando numa alteração da % de identidade nucleotídica, como é o caso de algumas sequências do coelho, da vaca e do cavalo - **Tabela 10** e **Figura E, F e G do Material suplementar**, respetivamente.

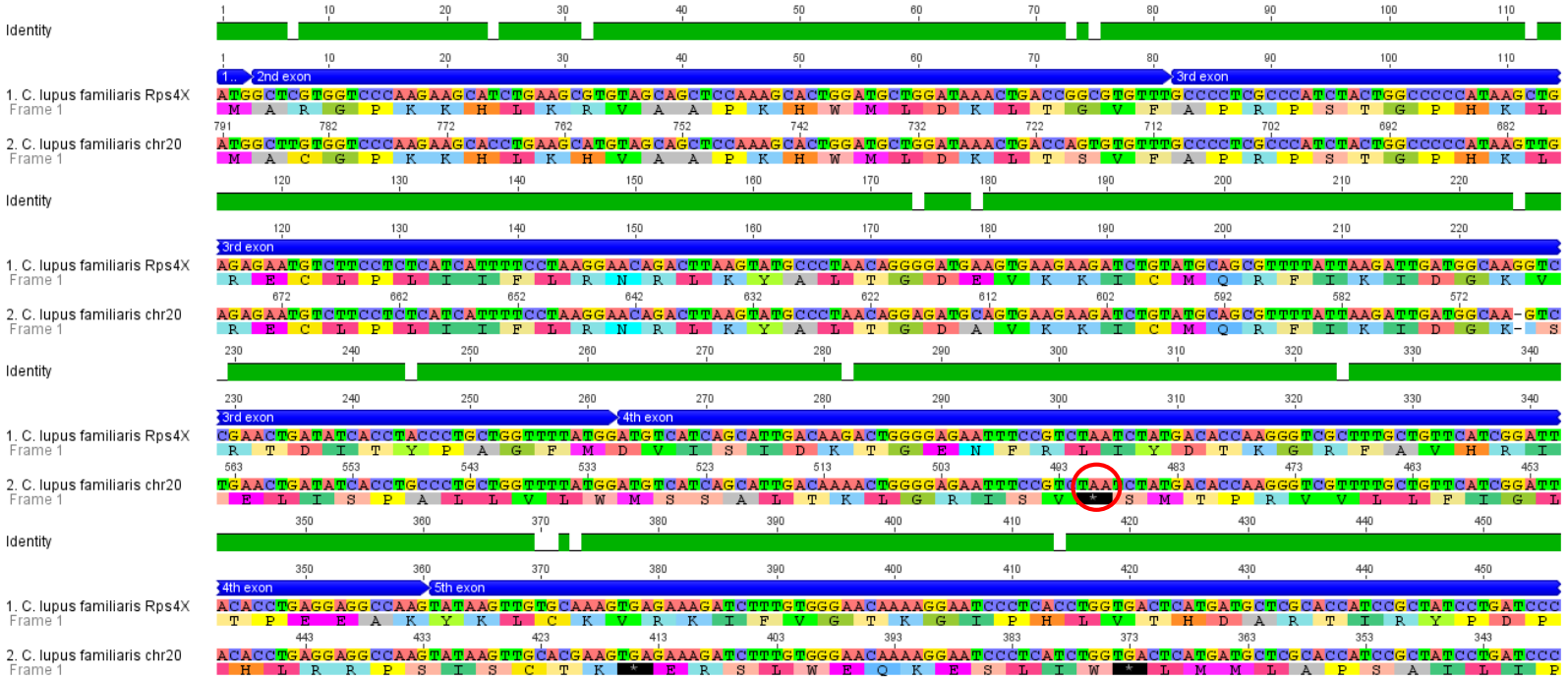


Figura 12 - Alinhamento e tradução *in silico* da sequência nucleotídica resultante do Blat com a *RPS4X* de *C. lupus familiaris* no genoma da própria espécie. Esta sequência do cromossoma 20 do cão é um exemplo com um codão STOP prematuro e que foi excluída da análise.Com o alinhamento confirma-se a posição inicial do "ATG" no exão 1 e, assinalado com um círculo vermelho, está o codão de terminação prematuro. A azul estão representados os exões da *RPS4X*.

Figura 13 - Alinhamento e tradução *in silico* da sequência nucleotídica do cromossoma 3 com a *Rps4x* de *M. musculus* no genoma da própria espécie. Esta sequência é um exemplo em que estava ausente o codão de iniciação (A), pretendendo-se demonstrar a pesquisa de sequência a 5' (acréscimo de 5bp relativamente à sequência encontrada por Blat) (B). Observando o alinhamento A, verifica-se que ao procurar o primeiro "ATG" na sequência resultante do Blat (marcado com um círculo vermelho) iria haver perda do primeiro exão e parte do segundo. Então procurou-se um codão de iniciação a montante (B), sendo encontrado o "ATG" do primeiro exão a 5' da sequência genómica identificada por Blat (marcado com um círculo vermelho), mas a ocorrência de algumas mutações resulta num codão de terminação prematuro e portanto esta sequência não prossegue para análise estrutural. A azul estão representados os exões anotados para a *Rps4x*.

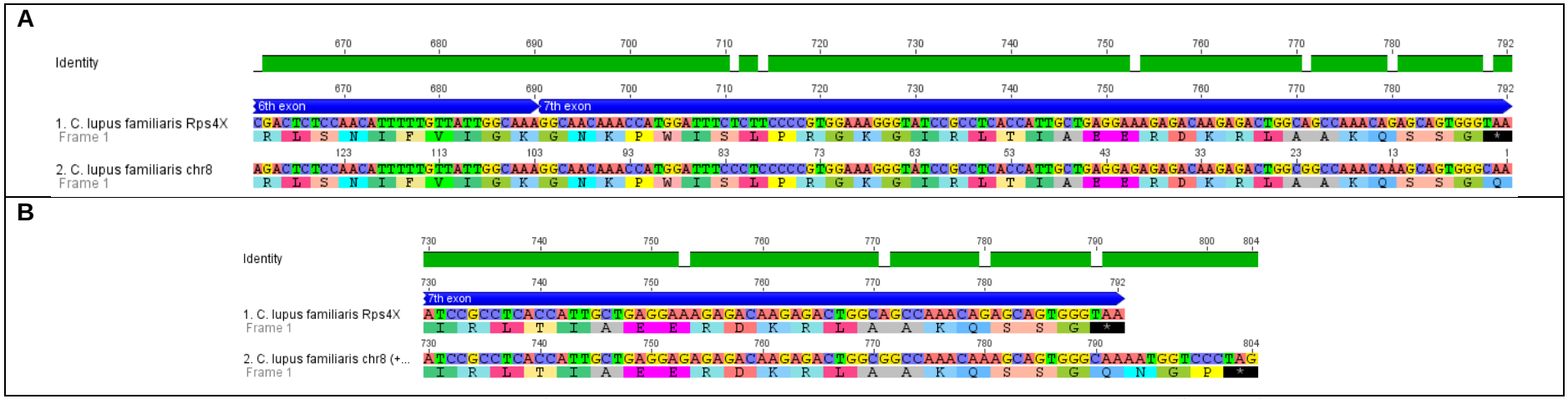


Figura 14 - Alinhamento e tradução *in silico* da sequência nucleotídica do cromossoma 8 com a *RPS4X* de *C. lupus familiaris* no genoma da própria espécie. A sequência apresentada em (A) exemplifica o caso em que existe sequência em falta a 3' porque, dada a mutação que ocorreu - c. 790 T>C - o codão STOP foi transformado num codão não-STOP, o qual codifica uma Glutamina e portanto, houve a necessidade de procurar o codão de terminação. Em (B) pretende-se demonstrar a pesquisa de sequência a 3' (acréscimo de 12bp relativamente à sequência encontrada por Blat). Esta sequência apresenta ORF e portanto é candidata a análise estrutural. A azul estão representados os exões anotados para a *RPS4X*.

2.2. Análise filogenética e evolutiva dos retrogenes da S4X

Para caraterizar as relações evolutivas entre os retrogenes encontrados na pesquisa anterior [da **Tabela 10**] e a S4X das diversas espécies construiu-se uma árvore filogenética com as sequências nucleotídicas (**Figura 15**) e outra com as sequências aminoacídicas (**Figura 16**). Neste estudo foram também incluídos *outgroups* (os mesmos que foram usados para a análise dos ortólogos da *Rps4l*).

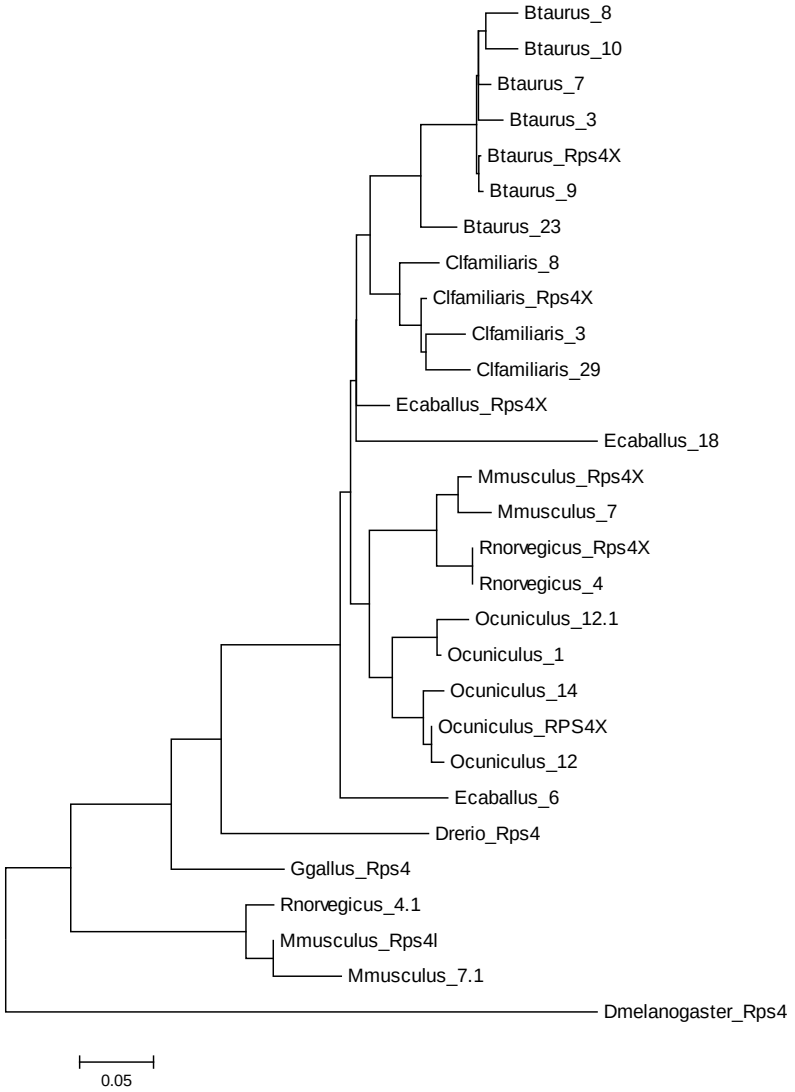


Figura 15 - Árvore filogenética do CDS da RPS4X e dos duplicados encontrados nos mamíferos estudados, obtida através do Mega6 com os parâmetros K2+G (Kimura 2-parameter + Gamma distribution). O *outgroup* mais divergente das sequências analisadas é a *Drosophila melanogaster*, uma vez que pertence à classe mais distante presente na árvore - *Insecta*. As espécies *Danio rerio* e *Gallus gallus*, mesmo sendo *outgroups*, apresentam sequências *RPS4* com maior identidade relativamente às analisadas, devido à proximidade das suas classes, *Peixes* e *Aves*, respetivamente - com os mamíferos.

Como esperado, os retrogenes selecionados agrupam maioritariamente por espécie, refletindo a árvore filogenética dos mamíferos (**Figura H do Material suplementar**) e a origem recente dos retrogenes a partir da RPS4X em cada linhagem. No entanto, existem também casos em que dois duplicados são mais semelhantes entre si do que com a RPS4X da mesma espécie (ex. *Btaurus*_10 e *Btaurus*_8), indicando que deverão representar eventos de duplicação secundários nessa espécie. Outros, por sua vez, divergiram consideravelmente, e agrupam com duplicados de espécies diferentes (ex. *Ecaballus*_6).

A sequência *Rnorvegicus*_4 é exatamente igual à RPS4X e portanto o seu estudo estrutural seria redundante e a sequência *Rnorvegicus*_4.1, correspondente à Rps4l desta espécie, já foi analisada no capítulo anterior juntamente com os restantes ortólogos dos roedores.

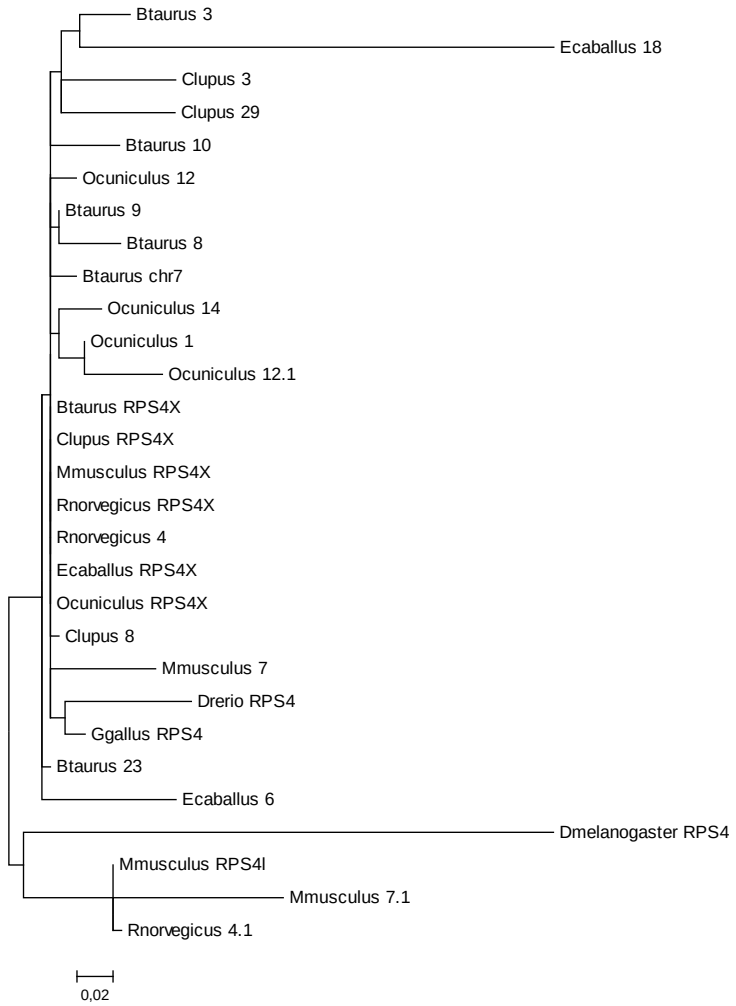


Figura 16 - Árvore filogenética das sequências aminoacídica da RPS4X e dos duplicados encontrados nos mamíferos estudados, obtida através do Mega6 com os parâmetros JTT+G (Jones-Taylor-Thornton+ Gamma distribution). O *outgroup* mais divergente das sequências analisadas é a *Drosophila melanogaster*, pelo disposto anteriormente na legenda da Figura 15. As espécies *Danio rerio* e *Gallus gallus*, mesmo sendo *outgroups*, apresentam sequências RPS4 com maior semelhança aminoacídica relativamente às analisadas, também pelo exposto na legenda da Figura 15.

As sequências proteicas da RPS4X agrupam todas juntas, uma vez que apesar das diferenças nucleotídicas, a nível aminoacídico todas as sequências são iguais, devido à redundância do código genético. As RPS4L de *M. musculus* e *R. norvegicus* (*Rnorvegicus*_4.1), como esperado, agrupam juntas, sendo o duplicado *Mmusculus*_7.1 o mais divergente, apesar de agrupar com as anteriores. Os duplicados dos mamíferos agrupam maioritariamente juntos e com as RPS4X mas existem casos, como as sequências *Btaurus*_23 e *Ecaballus*_6, em que se juntam num ramo diferente das anteriores, denotando divergência relativamente às mesmas. As sequências que divergiram menos relativamente à RPS4X e que, portanto, representam duplicados mais recentes ou mais conservados, foram a *Btaurus*_10, *Btaurus*_7, *Ocuniculus*_12 e *Clupus*_8. As sequências pertencentes ao murganho não foram analisadas uma vez que nesta espécie já é conhecida a presença de um retrogene da RPS4X – a RPS4L.

De modo a avaliar as pressões evolutivas que atuam sobre os retrogenes identificados em diferentes linhagens de mamíferos, calculou-se a razão entre as substituições sinónimas e não sinónimas (dN/dS) entre cada retrogene e o gene ancestral do cromossoma X de cada espécie (**Tabela 11**). As substituições sinónimas não levam à alteração do gene, e são neutras, enquanto que as não sinónimas estão sujeitas à ação da seleção.

Tabela 11 - Taxas de substituições sinónimas e não sinónimas entre os retrogenes analisados em mamíferos e respetiva RPS4X da mesma espécie.

Espécie	N	S	dN	dS	dN/dS
***O. cuniculus* chr 1**	574,9	214,1	0,0071	0,2059	0,0345
***O. cuniculus* chr 14**	563,5	225,5	0,0147	0,0324	0,4533
***C. l. familiaris* chr 8**	596,8	192,2	0,0034	0,1910	0,0178
***B. taurus* chr 23**	567,2	221,8	0,0036	0,2428	0,0148
***B. taurus* chr 9**	527,8	258,2	0,0039	0,0039	1,0154
***B. taurus* chr 7**	563,5	225,5	0,0073	0,0226	0,3219
***E. caballus* chr 6**	570,9	218,1	0,0406	0,2683	0,1513

Resultados obtidos com o PAL2NAL. N - número de sítios não sinónimos; S - número de sítios sinónimos; dN - taxa de substituições não sinónimas; dS - taxa de substituições sinónimas.

Também nos mamíferos analisados, a relação dN/dS entre a RPS4X e os respetivos duplicados em cada espécie apresenta valores muito baixos, inferiores à unidade, com a exceção do retrogene do cromossoma 9 da vaca. Relembrando, os valores da relação dN/dS inferiores à unidade indicam um maior número de substituições sinónimas do que não sinónimas desde que ocorreu a duplicação. Ou seja, as pressões evolutivas que têm atuado sobre as sequências visam a conservação do seu estado ancestral - seleção negativa. Quando a relação dN/dS é aproximadamente igual à unidade (cromossoma 9

da vaca), ocorreu um número igual de substituições sinónimas e não sinónimas desde a duplicação, o que indica ausência de pressão seletiva para manter a função do gene e deverá refletir-se em pseudogenização.

2.3. Análise estrutural dos retrogenes da S4X

2.3.1 Seleção de duplicados para análise estrutural

Para restringir o estudo aos duplicados com maior potencial de manterem conservada a função da RPS4X, fez-se a tradução *in silico* das sequências selecionadas seguindo os critérios anteriormente descritos e para cada espécie fez-se o alinhamento dos duplicados com a respetiva RPS4X (**Figura 17, 20, 23 e 26**), sendo apenas analisadas as que apresentavam uma similaridade aminoacídica superior a 90% (**Tabela 10**). Seguidamente fez-se uma previsão *in silico* com o programa **SIFT** do impacto das substituições encontradas na função da proteína, classificando-as em toleradas ou não-toleradas (**Tabela IV, V, VI E VII do Material suplementar**), podendo estas alterações estar em posições conservadas ou não conservadas do alinhamento da RPS4, como descrito no capítulo anterior para a S4L (**Figura 7**).

Alguns dos duplicados identificados apresentam muitas alterações aminoacídicas e será mais improvável que estes mantenham a mesma função da RPS4X. Assim, foram definidos critérios de forma a selecionar para a análise estrutural apenas os duplicados com maior probabilidade de poderem compensar a função da S4X da mesma espécie. Dado que a presença de várias alterações aminoacídicas no mesmo domínio funcional aumenta a probabilidade de destabilização da estrutura terciária e de que a função desse domínio fique comprometido, somente foram consideradas as sequências que apresentavam apenas uma alteração não tolerada em cada domínio crucial para a função da proteína (*N-terminal* e *RNA-binding*) e no máximo duas nos restantes domínios. As alterações fora dos domínios proteicos não foram consideradas na análise estrutural, dada a sua menor importância a nível funcional. As sequências resultantes desta filtragem estão na **Tabela 12**.

Tabela 12 - Duplicados que foram selecionados para análise estrutural e respetivo número de alterações não toleradas em posições conservadas relativamente à RPS4X da respetiva espécie.

Espécie	Cópias autossómicas	Número de alterações não toleradas dentro dos domínios
Oryctolagus cuniculus	Chr 1	1
	Chr14	1
Canis lupus familiaris	Chr8	1
Bos taurus	Chr23	1
	Chr 9	0
	Chr 7	0
Equus caballus	Chr6	5

A sequência do cromossoma 6 de cavalo, apesar de não se enquadrar nos limites definidos devido a ter 5 alterações em posições conservadas (2 no domínio *N-terminal* e 3 no domínio *RNA-binding)*, foi analisada porque era a única disponível nesta espécie.

Ambos os programas são coerentes quanto aos resultados, ou seja, as posições não toleradas pelo **SIFT** (**Tabela IV, V, VI e VII do Material suplementar**) são também identificadas como deletérias pelo **PROVEAN Protein** (**Tabela 13**), uma vez que os dois se baseiam no alinhamento de sequências com elevada semelhança com a sequência em questão, para prever o impacto das substituições aminoacídicas.

Tabela 13 - Resultados do PROVEAN Protein para as posições não toleradas detetadas pelo SIFT dentro dos domínios e que ocorrem em locais conservados da RPS4 (Figura 7).

Espécie e duplicado	Posição	PROVEAN Protein
O. cuniculus **chr 1 e 14**	25	Deletéria
C. l. familiaris **chr 8**	32	Deletéria
B.taurus **chr 23**	95	Deletéria
E. caballus **chr 6**	31	Deletéria
	34	Deletéria
	44	Deletéria
	55	Deletéria
	67	Deletéria

2.3.2. Análise estrutural dos retrogenes S4X

Para avaliar o impacto das diferenças aminoacídicas entre a S4X e os seus retrogenes na função das proteínas codificadas por estes modelou-se a estrutura tridimensional de cada um utilizando como molde o modelo de RPS4 de *O. Cuniculus* (4KZX)[58], obtido a partir de ribossomas de reticulócitos, tal como se procedeu na análise da S4L.

Oryctolagus cuniculus

Observando o alinhamento da **Figura 17**, a sequência identificada como 12.1 é, claramente, a que apresenta mais alterações aminoacídicas, essencialmente na parte terminal, já fora dos domínios, devido a uma deleção *out of frame* de um nucleótido (**Figura E do Material suplementar**) tendo um acréscimo de 5 aminoácidos (269 aa) relativamente à RPS4X e aos outros duplicados (264 aa) desta espécie.

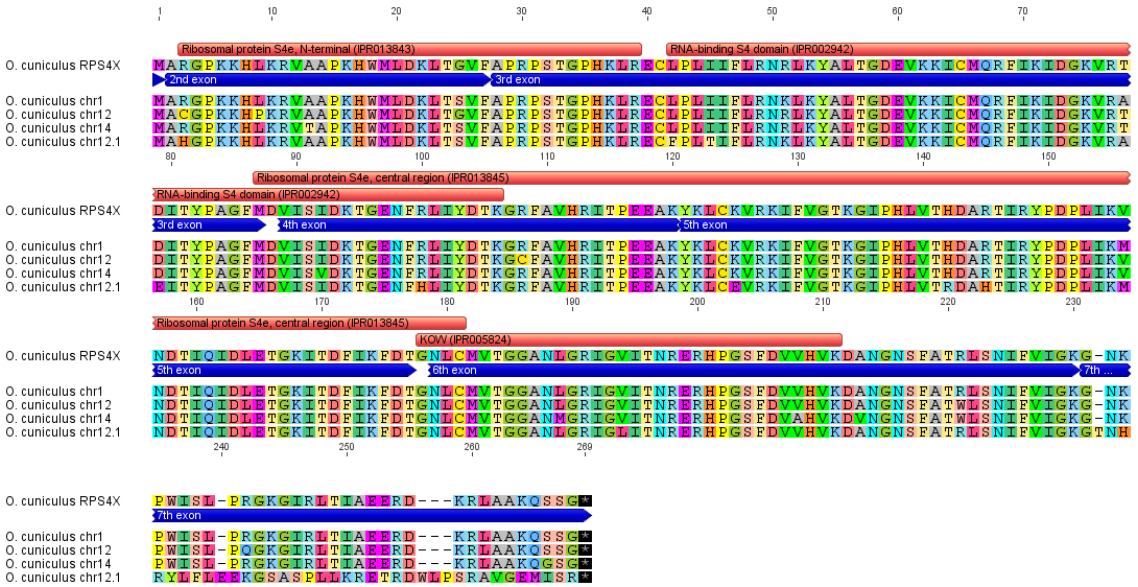


Figura 17 - Alinhamento da sequência aminoacídica obtida por tradução *in silico* dos duplicados do coelho selecionados para análise funcional. Existem dois duplicados localizados no cromossoma 12, que foram identificados como 12 e 12.1.

Devido ao elevado número de alterações aminoacídicas em regiões conservadas e à sua grande divergência relativamente à RPS4X (**Figura 17** e **Tabela IV do Material suplementar**), a sequência 12.1 não foi considerada para análise estrutural. Também a sequência 12 foi excluída de análises posteriores uma vez que não se enquadrava nos limites previamente expostos. Apenas as sequências do cromossoma 1 e 14 foram consideradas.

Nos duplicados selecionados para análise estrutural figura apenas uma alteração aminoacídica localizada no domínio *N-terminal* (**Tabela 14**), em que a glicina na posição 25 da RPS4X foi substituída por uma serina. Pelo facto de ser uma substituição não tolerada, foram estudadas as interações polares para essa posição em ambas as sequências, de for.ma a averiguar se as ligações se mantêm ou se há modificações (**Figura 18**).

Tabela 14 – Substituição aminoacídica não tolerada nos duplicados de coelho.

Substituições não toleradas entre a RPS4X de coelho e os seus duplicados						
Domínio	Posição	C/NC relativamente à RPS4	RPS4X	Chr1	Chr14	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	25	C	G	S	S	Y G

Apenas uma posição conservada da RPS4 (Figura 7) apresenta uma substituição não tolerada, a qual é comum aos dois duplicados que foram analisados nesta espécie (cromossoma 1 e 14). *As letras maiúsculas indicam aminoácidos que existem nessa posição em outras espécies e portanto aparecem no alinhamento realizado pelo SIFT.

Verifica-se então, que a glicina-25 na RPS4X não estabelece interações polares com outros aminoácidos (**Figura 18A**) ao passo que a serina-25 estabelece interações polares com a treonina-24 e a valina-26 (**Figura 18B**).

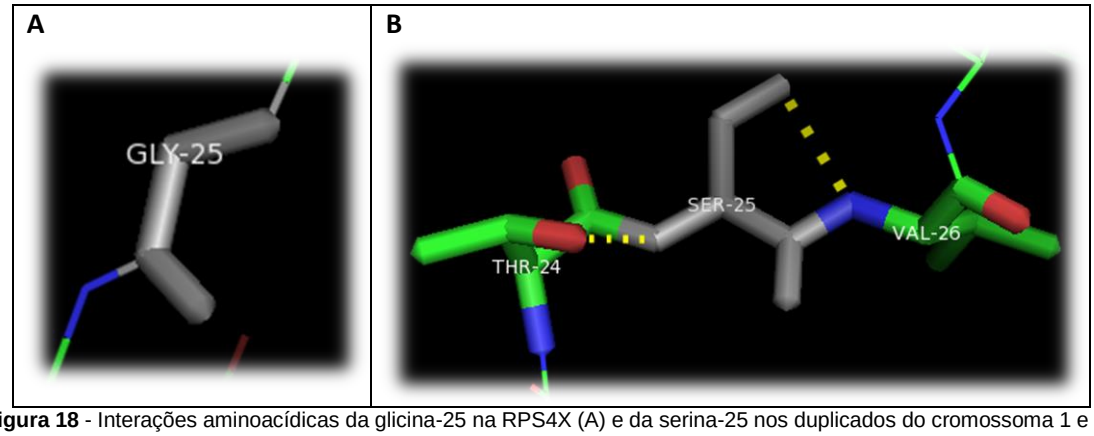


Figura 18 - Interações aminoacídicas da glicina-25 na RPS4X (A) e da serina-25 nos duplicados do cromossoma 1 e 14 (B) de coelho. Uma vez que o aminoácido alterado e as interações polares são iguais em ambos os duplicados, apresenta-se apenas uma imagem (B). A cinza está representado o aminoácido alterado entre as duas sequências e a amarelo tracejado estão representadas as interações polares entre aminoácidos, as quais diferem entre a RPS4X e os duplicados.

Com o intuito de confirmar se a alteração descrita na **Figura 18** para a posição 25 poderá causar modificações a nível estrutural, fez-se a sobreposição das estruturas da RPS4X e dos duplicados desta espécie, bem como das respetivas interações. Como os duplicados são diferentes, fez-se a sua sobreposição com a RPS4X, separadamente (**Figura 19A e Figura 19B**). Observando a **Figura 19C**, tendo em conta a baixa resolução da estrutura modelo disponível, não são aparentes alterações estruturais em nenhum dos duplicados e, portanto, pensa-se que esta modificação não terá influência do ponto de vista funcional nestas sequências.

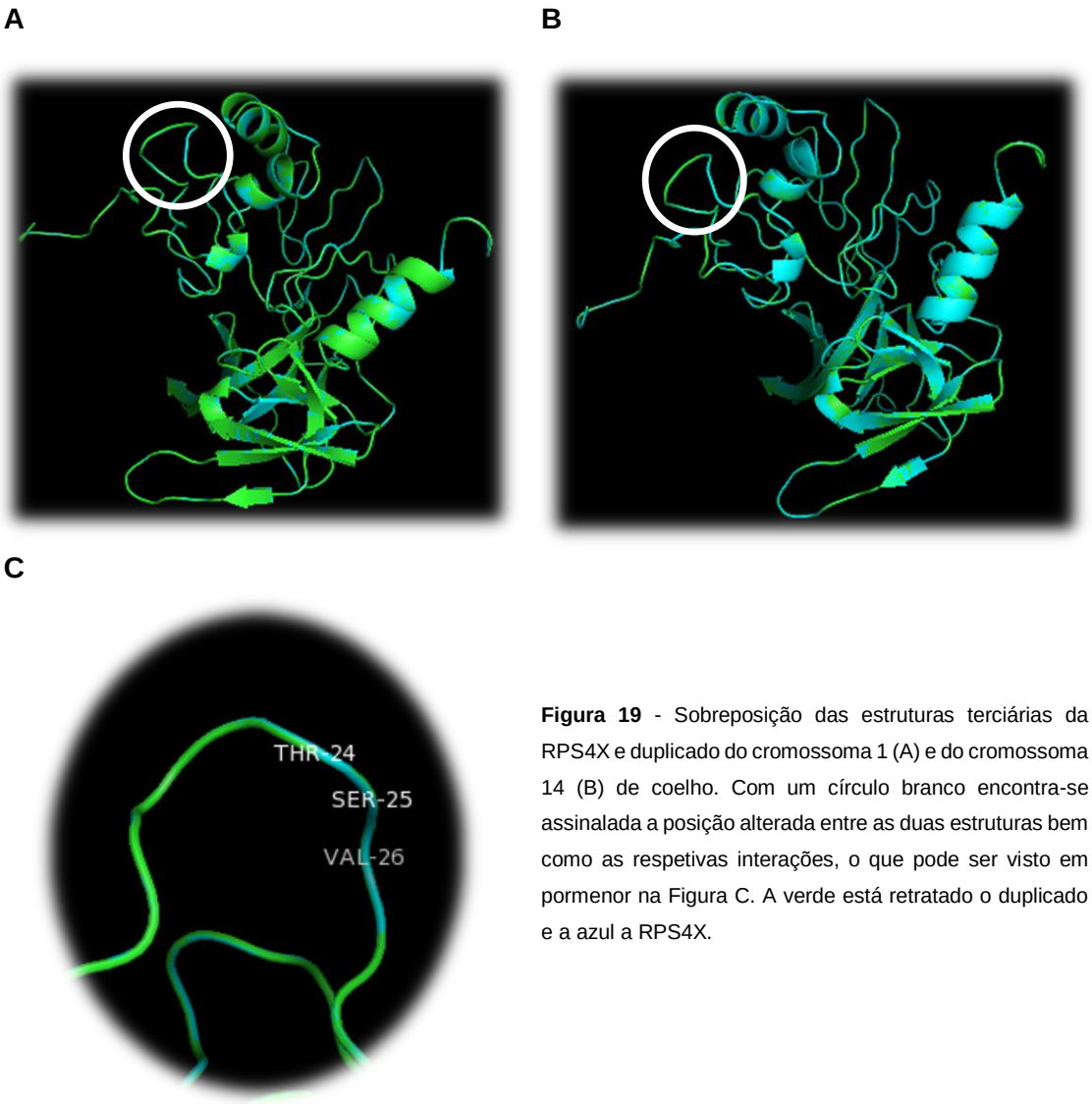


Figura 19 - Sobreposição das estruturas terciárias da RPS4X e duplicado do cromossoma 1 (A) e do cromossoma 14 (B) de coelho. Com um círculo branco encontra-se assinalada a posição alterada entre as duas estruturas bem como as respetivas interações, o que pode ser visto em pormenor na Figura C. A verde está retratado o duplicado e a azul a RPS4X.

Canis lupus familiaris

Observando o alinhamento da **Figura 20**, verifica-se que as sequências dos cromossomas 3 e 29 são mais curtas (devido à presença de um STOP prematuro - 221 aa) em relação à RPS4X, ao passo que a do cromossoma 8 é mais longa (268 aa) pelo exposto na legenda da **Figura 14**.

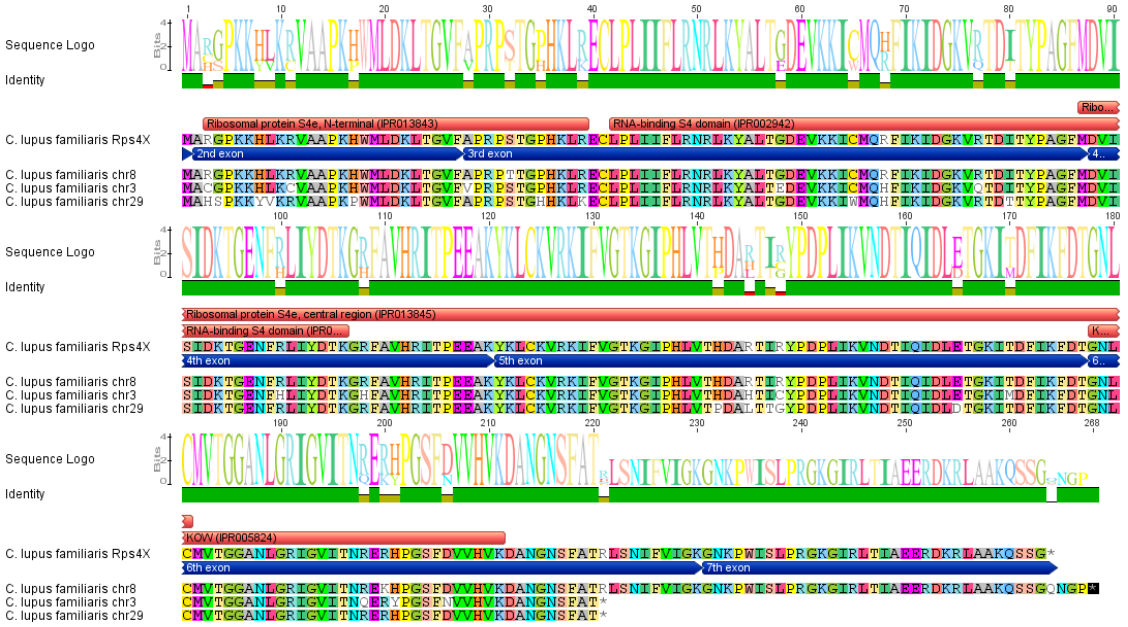


Figura 20 - Alinhamento da sequência proteica obtida por tradução *in silico* dos duplicados selecionadas para a espécie *C. lupus familiaris*.

Devido à elevada quantidade de alterações aminoacídicas em posições conservadas (**Figura 20 e Tabela V do Material suplementar**), tanto a sequência do cromossoma 3 como a do 9 foram excluídas da análise estrutural, sendo apenas a sequência do cromossoma 8 candidata a essa análise.

Neste duplicado figura apenas uma alteração aminoacídica localizada no domínio *N-terminal* (**Tabela 15**), em que a serina na posição 32 da RPS4X foi substituída por uma treonina no duplicado.

Sendo esta substituição não tolerada, estudaram-se as interações polares para essa posição em ambas as sequências, de forma a averiguar se as ligações se mantêm ou se há modificações (**Figura 21**).

Tabela 15 - Alteração aminoacídica não tolerada, no duplicado do cromossoma 8 de cão, numa posição conservada da RPS4.

Substituições não toleradas entre a RPS4X do cão e os seus duplicados					
Domínio	Posição	C/NC relativamente à RPS4	RPS4X	Chr8	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	32	C	S	T	N A S

*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT.

Verifica-se então que na RPS4X a serina-32 estabelece duas ligações polares com a treonina-81 (**Figura 21A**), ao passo que no duplicado a treonina-32 estabelece duas interações polares com a treonina-33 e uma com a treonina-81 (**Figura 21B**).

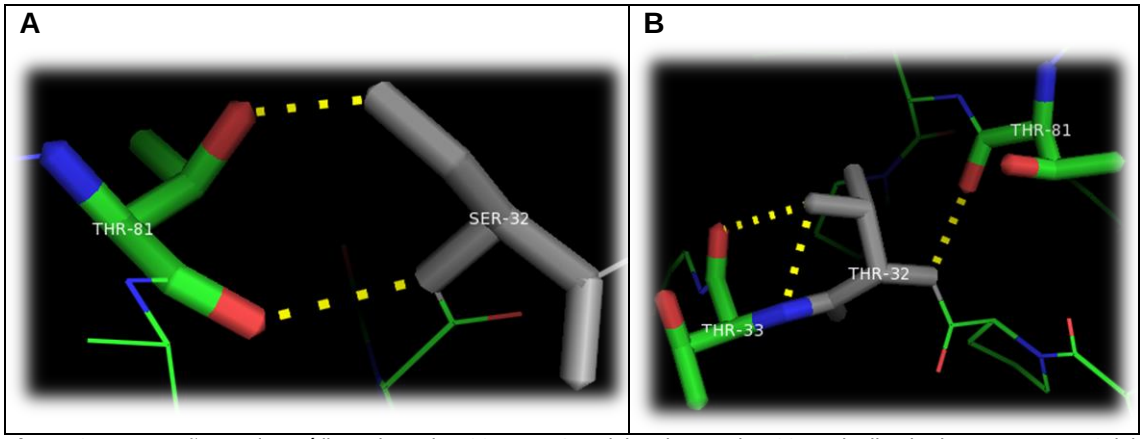


Figura 21 - Interações aminoacídicas da serina-32 na RPS4X (A) e da treonina-32 no duplicado do cromossoma 8 (B) de cão. A cinza está representado o aminoácido alterado entre as duas sequências e a amarelo tracejado estão representadas as interações polares entre aminoácidos, as quais diferem entre a RPS4X e o duplicado.

Com o intuito de confirmar se a alteração descrita na **Figura 21** para a posição 32 poderá causar modificações a nível estrutural, fez-se a sobreposição das estruturas terciárias da RPS4X e do duplicado desta espécie, bem como das respetivas interações. Observando a **Figura 22** não são aparentes alterações, portanto, pensa-se que esta modificação não deverá ter influência a nível funcional.

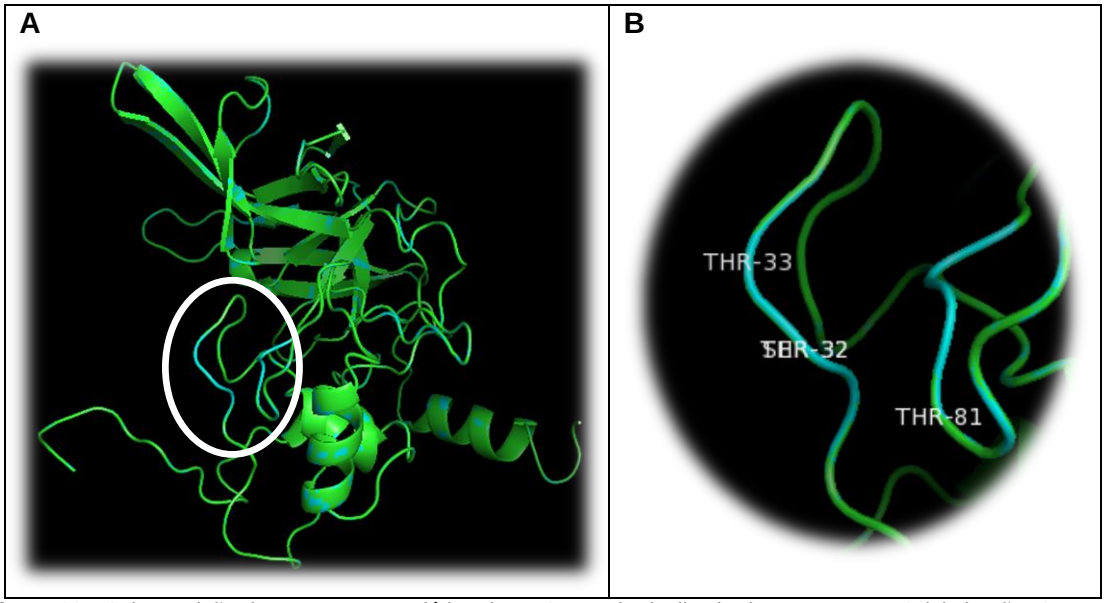


Figura 22 - Sobreposição das estruturas terciárias da RPS4X e do duplicado do cromossoma 8 (A) de cão. Com um círculo branco encontra-se assinalada a posição alterada entre as duas estruturas bem como as respetivas interações, o que pode ser visto em pormenor na Figura B. A verde está retratado o duplicado e a azul a RPS4X.

Bos taurus

Com base no alinhamento da **Figura 23**, confirma-se que a sequência do cromossoma 3 é a mais curta (250 aa) relativamente à RPS4X, pois apresenta um codão STOP prematuro devido à inserção de um nucleótido que altera a grelha de leitura **(Figura F do Material suplementar)**. Também a sequência do cromossoma 9 tem um aminoácido a menos (263 aa) enquanto os outros duplicados se apresentam com 264 aa.

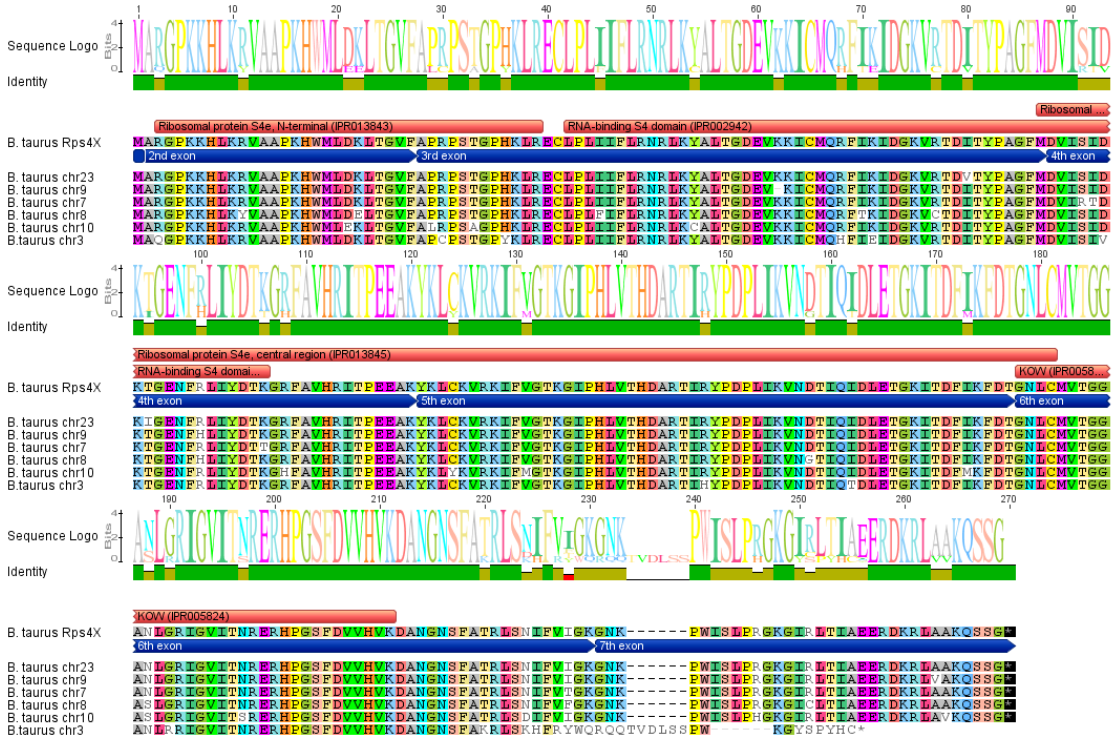


Figura 23 - Alinhamento da sequência aminoacídica obtida por tradução *in silico* dos duplicados selecionados para a espécie *B. taurus*.

Devido à elevada quantidade de alterações aminoacídicas em posições conservadas (**Figura 23 e Tabela VI do Material suplementar**), as sequências dos cromossomas 8 e 3 foram excluídas da análise estrutural. Também a sequência do cromossoma 10 foi excluída da análise posterior, uma vez que não se enquadrava nos critérios definidos anteriormente. Para a análise estrutural seguiram as sequências dos cromossomas 23, 9 e 7. As sequências 9 e 7 não apresentam alterações aminoacídicas não toleradas em posições conservadas da RPS4 (**Figura 7**), daí não constarem na **Tabela 16**.

No duplicado do cromossoma 23 existe apenas uma alteração aminoacídica localizada no domínio *RNA-binding* (**Tabela 16**), em que a treonina na posição 95 da RPS4X foi substituída por uma isoleucina no duplicado. Sendo esta substituição não tolerada,

estudaram-se as interações polares para essa posição em ambas as sequências, de forma a averiguar se as ligações se mantêm ou se há modificações (**Figura 24**).

Tabela 16 - Alteração aminoacídica não tolerada, no duplicado do cromossoma 23 da vaca numa posição conservada da RPS4 (Figura 7).

Posições não toleradas entre a RPS4X da vaca e os seus duplicados					
Domínio	Posição	C/NC relativamente à RPS4	RPS4X	Chr23	Aminoácidos tolerados*
RNA-binding S4 (IPR002942) S4e, central region (IPR013845)	95	C	T	I	T

Existe apenas uma alteração aminoacídica onde há a substituição de uma treonina por uma isoleucina. *As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT.

Verifica-se então, que a treonina-95 na RPS4X estabelece interações polares com a isoleucina-92, ácido aspártico-93, glicina-96 e ácido glutâmico-97 (**Figura 24A**) ao passo que, a isoleucina-95 no duplicado estabelece apenas uma ligação com o ácido aspártico-93 (**Figura 24B**).

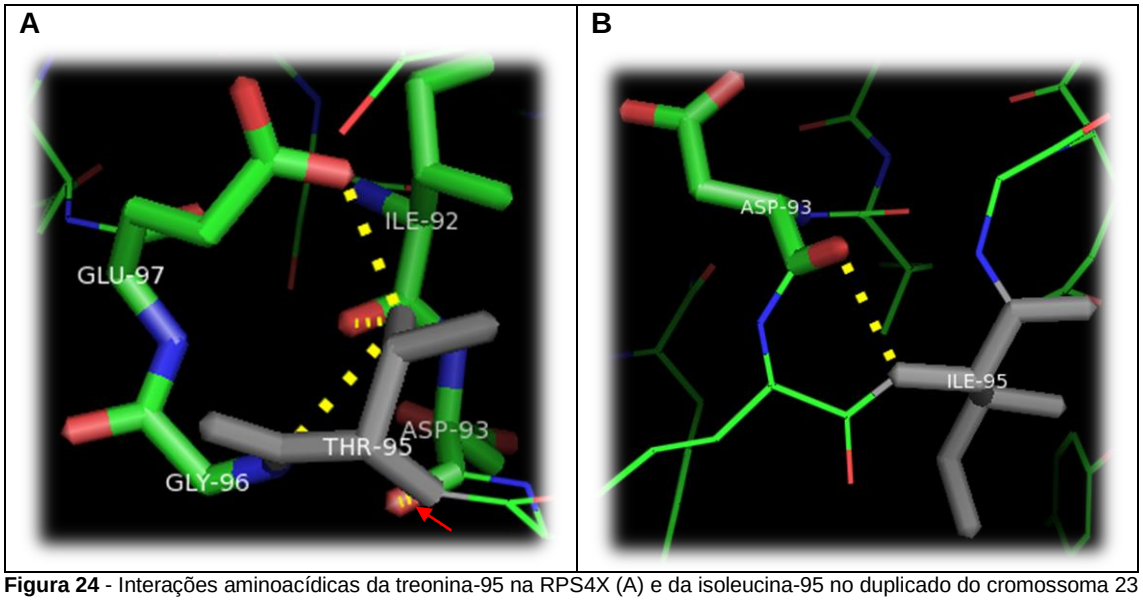


Figura 24 - Interações aminoacídicas da treonina-95 na RPS4X (A) e da isoleucina-95 no duplicado do cromossoma 23 (B) desta espécie. A cinza está representado o aminoácido alterado entre as duas sequências e a amarelo tracejado estão representadas as interações polares entre aminoácidos (uma das ligações está assinalada com uma seta vermelha, devido a ser menos percetível), as quais diferem entre a RPS4X e o duplicado.

Com o intuito de confirmar se a alteração descrita na **Figura 24** para a posição 95 poderá causar modificações a nível estrutural, fez-se a sobreposição das estruturas terciárias da RPS4X e do duplicado da espécie em questão, bem como das respetivas

interações. Observando a **Figura 25**, a nível da estrutura global não são aparentes alterações, portanto não se prevê que esta modificação tenha influência a nível funcional.

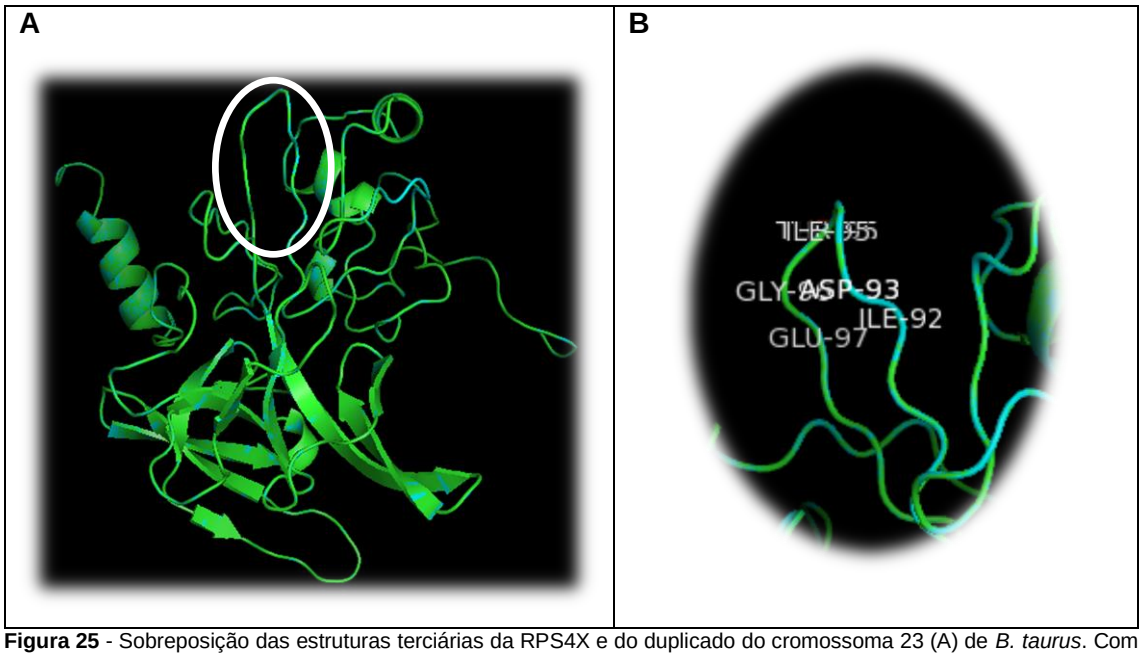


Figura 25 - Sobreposição das estruturas terciárias da RPS4X e do duplicado do cromossoma 23 (A) de *B. taurus*. Com um círculo branco encontra-se assinalada a posição alterada entre as duas estruturas bem como as respetivas interações, o que pode ser visto a pormenor na Figura B. A verde está retratado o duplicado e a azul a RPS4X.

Equus caballus

Observando o alinhamento da **Figura 26**, a sequência do cromossoma 18 é mais curta (222 aa) relativamente à RPS4X e à sequência do cromossoma 6 (264 aa) pois apresenta um codão STOP prematuro (**Figura G do Material suplementar).**

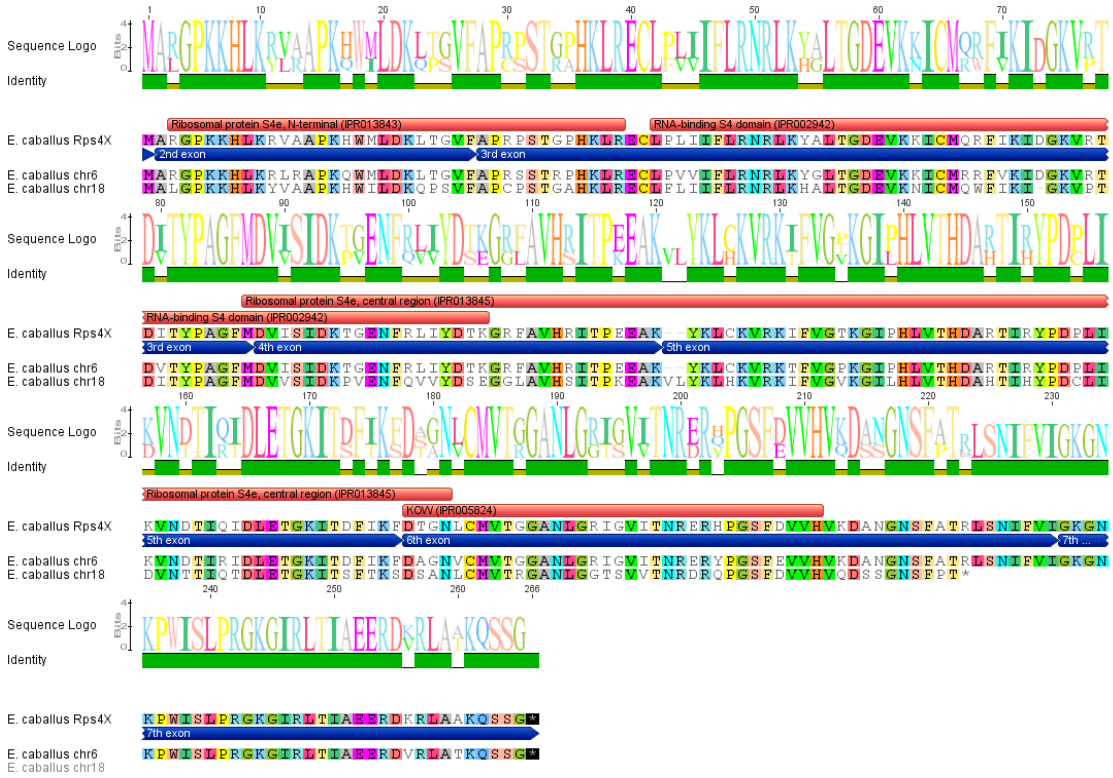


Figura 26 - Alinhamento da sequência aminoacídica obtida por tradução *in silico* dos duplicados selecionados para a espécie *E. caballus.*

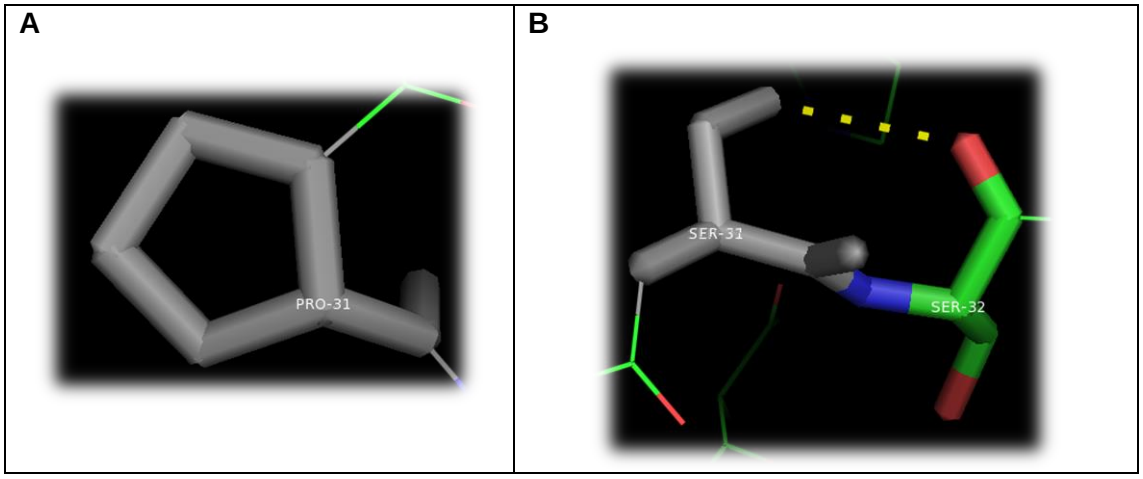
A sequência do cromossoma 18 foi excluída da análise estrutural devido ao elevado número de alterações aminoacídicas em posições conservadas (**Figura 26 e Tabela VII do Material suplementar**). Apesar da sequência do cromossoma 6 apresentar duas substituições no domínio *N-terminal* e 3 no *RNA-binding* (**Tabela 17**), decidimos analisá-la ao nível da estrutura, já que dos duplicados que encontramos este tem maior potencial para manter a função da RPS4X nesta espécie.

Tabela 17 - Alterações aminoacídicas não toleradas no duplicado do cromossoma 6 de cavalo, em posições conservadas da RPS4 (Figura 7).

Substituições não toleradas entre a RPS4X de cavalo e os seus duplicados					
Domínio	Posição	C/NC relativamente à RPS4	RPS4X	Chr6	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	31	C	P	S	P
	34	C	G	R	G
RNA-binding S4 (IPR002942)	44	C	L	V	L
	55	C	A	G	A
	67	C	Q	R	E Q

Existem cinco substituições aminoacídicas entre a RPS4X e o duplicado analisado. *As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT.

A prolina-31 (**Figura 27A**) e a glicina-34 (**Figura 27C**) da RPS4X não apresentam interações polares com outros aminoácidos, ao passo que no duplicado a serina-31 (**Figura 27B**) e a arginina-34 (**Figura 27D**) apresentam interações com a serina-32 e o ácido aspártico-79, respetivamente. Na posição 44, a leucina na RPS4X (**Figura 27E**) e a valina no duplicado (**Figura 27F**) apresentam as mesmas interações polares com a fenilalanina-47, leucina-48 e tirosina-82. Também na posição 55, tanto a alanina na RPS4X (**Figura 27G**) como a glicina no duplicado (**Figura 27H**), apresentam uma interação polar com a leucina-48. Para a posição 67, a glutamina na RPS4X estabelece uma interação polar com a isoleucina-64 (**Figura 27I**) ao passo que, no duplicado, a arginina-67 estabelece duas interações polares, uma com a lisina-63 e outra com a isoleucina-64 (**Figura 27J**).



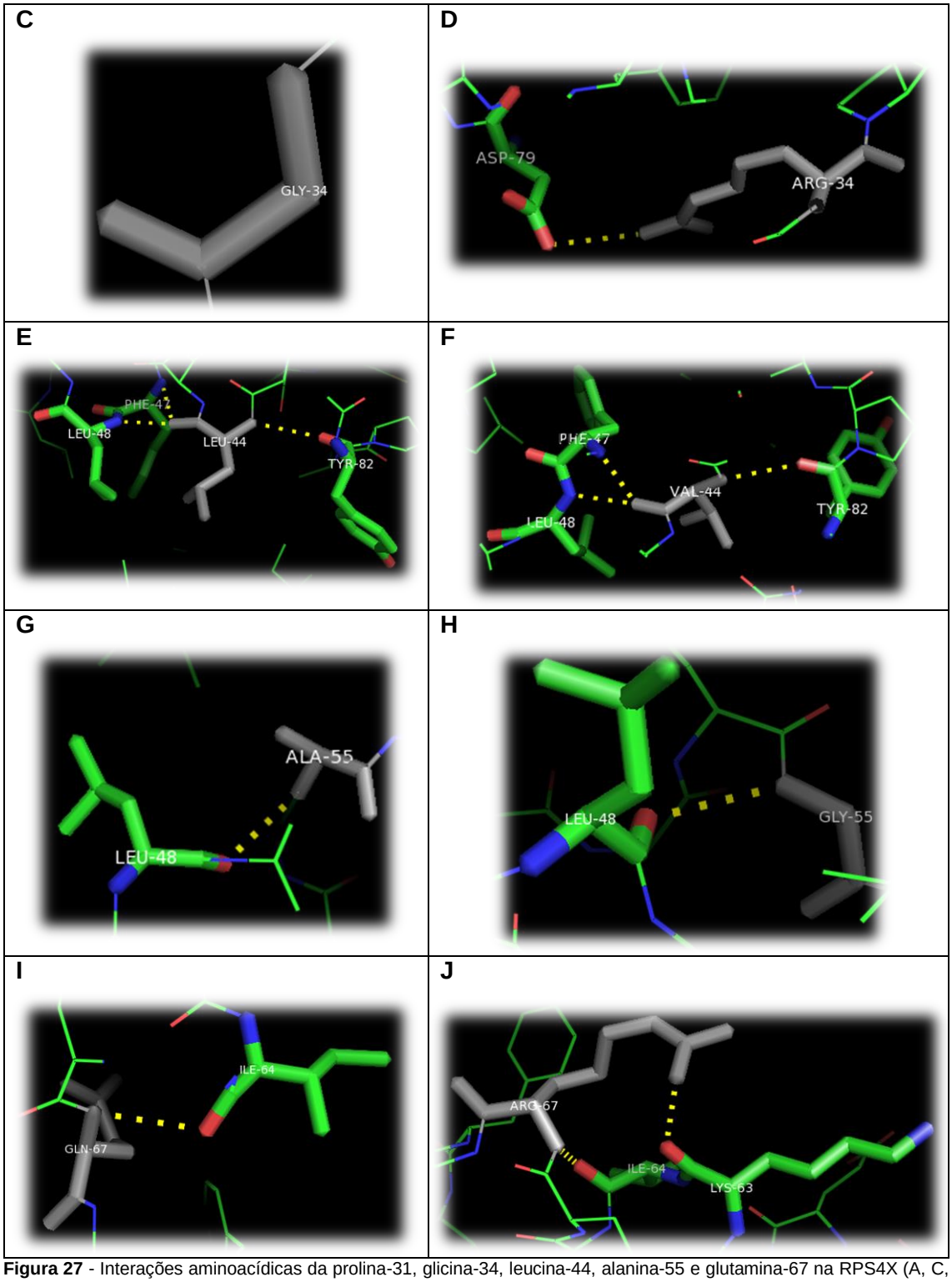


Figura 27 - Interações aminoacídicas da prolina-31, glicina-34, leucina-44, alanina-55 e glutamina-67 na RPS4X (A, C, E, G e I) e da serina-31, arginina-34, valina-44, glicina55 e arginina 67 no duplicado do cromossoma 6 (B, D, F, H e J) de cavalo. A cinza está representado o aminoácido alterado entre as duas sequências e a amarelo tracejado estão representadas as interações polares entre aminoácidos, as quais diferem entre a RPS4X e o duplicado.

Como a substituição dos aminoácidos para as posições 44 (**Figura 27E e Figura 27F**) e 55 (**Figura 27G e Figura 27H**) não altera as interações polares no duplicado relativamente à RPS4X, não se preveem modificações a nível estrutural para estas

posições. Para as restantes posições, 31, 34 e 67 fez-se a sobreposição das estruturas terciárias da RPS4X e do duplicado bem como das respetivas interações (**Figura 28**), mas não se observaram alterações nos respetivos elementos estruturais contendo essas alterações. No entanto, observando a **Figura 28A**, verifica-se que existe uma *β-sheet* mais curta no duplicado (assinalada com uma seta a vermelho) relativamente à RPS4X, o que poderá ser devido a outras alterações aminoacídicas no duplicado. Na **Figura 28B e 28C**, a nível estrutural não se detetam outras alterações.

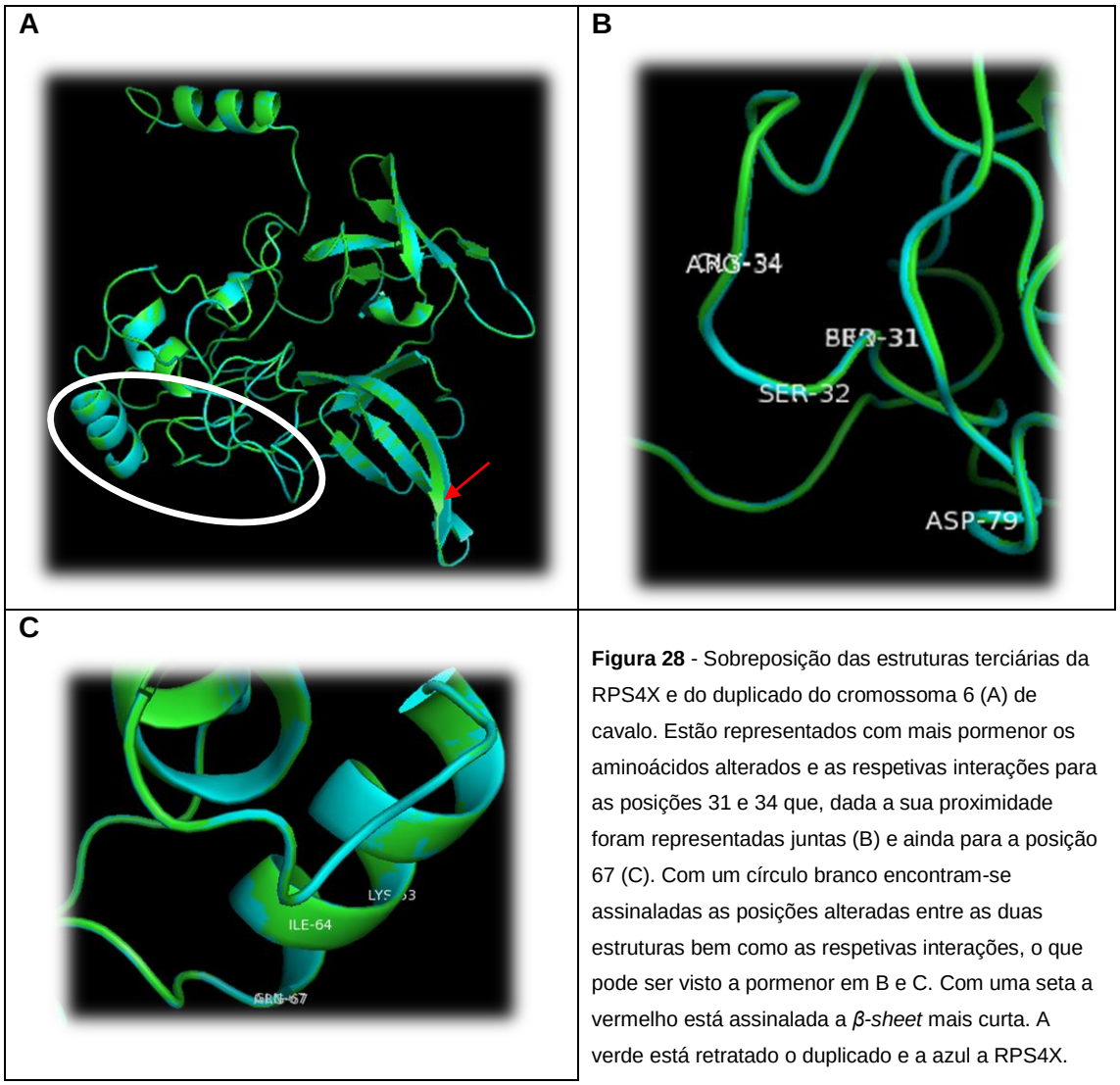


Figura 28 - Sobreposição das estruturas terciárias da RPS4X e do duplicado do cromossoma 6 (A) de cavalo. Estão representados com mais pormenor os aminoácidos alterados e as respetivas interações para as posições 31 e 34 que, dada a sua proximidade foram representadas juntas (B) e ainda para a posição 67 (C). Com um círculo branco encontram-se assinaladas as posições alteradas entre as duas estruturas bem como as respetivas interações, o que pode ser visto a pormenor em B e C. Com uma seta a vermelho está assinalada a *β-sheet* mais curta. A verde está retratado o duplicado e a azul a RPS4X.

Em resumo dos resultados obtidos e também dos resultados já existentes quanto a retrogenes da RPS4X, construiu-se a **Tabela 18**.

Tabela 18 - Resumo da informação sobre a presença ou ausência de cópia da RPS4 ligada ao Y e retrogenes encontrados com potencial funcional.

Espécie	RPS4Y	Retrogene
M. musculus	Não[9]	RPS4L[38]
R. norvegicus	Não[9]	RPS4[60]
O. cuniculus	?	Chr 1 e 14
C. l. familiaris	Não[35]	Chr 8
F. catus	2 cópias[35]	Não
B. taurus	Não[9]	Chr 23, 9 e 7
S. scrofa	?	?
E. caballus	Não[61]	Chr 6
H. sapiens	RPS4Y1/RPS4Y2[7,32]	Não

Discussão

Inicialmente os estudos da filogenia da *RPS4* sugeriam que a transposição do gene para o cromossoma X teria ocorrido antes da radiação dos mamíferos e que existiria uma cópia no cromossoma Y apenas em primatas.[33,62] Contudo, a descoberta do gene *RPS4Y* em marsupiais (*M. domestica*)[31] bem como a sua presença no bloco *X-degenerate* na espécie humana[7], refutou a especificidade da sequência nesta linhagem. Entretanto, também se verificou que algumas espécies de mamíferos não possuem cópia da *RPS4Y* (murganho[9], rato[9], cão[35], vaca[9] e cavalo[61]), que terá sido perdida nestas linhagens[2]. Em humanos existem 2 cópias do gene ligado ao cromossoma Y, sendo uma delas, *RPS4Y2*, expressa especificamente em testículo, em células da linha germinal[37]. Durante a inativação meiótica dos cromossomas sexuais na espermatogénese, estes apresentam uma atividade transcricional muito baixa. No entanto, após a conclusão da meiose, em murganho, existem pelo menos dois genes ligados ao Y (*Zfy* e *Sry)*, e dois genes do cromossoma X (*Ube1x* e *Ube2a*) que estão a ser transcritos em espermatídeos haploides, e portanto, são reativados[63]. No entanto não existem evidências que o gene *Rps4x* seja reativado após a IMCS nas espécies estudadas. Em murganho (*M. musculus*), em que não existe o gene *RPS4Y*, foi descrito um gene autossómico originado por retrotransposição, designado *Rps4l*, cuja proteína é expressa especificamente em testículo, em células da linha germinal e incorporada em polissomas neste tecido[38]. Estes resultados sugerem que a RPS4 será importante para a espermatogénese e que as cópias ligadas ao cromossoma Y ou aos autossomas poderão compensar o silenciamento da cópia do gene ligada ao X durante a inativação meiótica dos cromossomas sexuais, que ocorre na meiose masculina.

Assim, o presente trabalho consistiu na pesquisa de retrogenes autossómicos da *RPS4X* potencialmente funcionais em linhagens de mamíferos, que possam compensar o silenciamento da cópia ligada ao X no sexo heterogamético. Partimos da hipótese de que nas espécies em que ocorreu a perda do gene *RPS4* ligado ao cromossoma Y a retrotransposição da S4 para os autossomas terá sido selecionada positivamente e os retrogenes terão sido mantidos. E, contrariamente, quando a espécie possui genes *S4* ligados ao cromossoma Y, que poderão ser expressos após a meiose, poderão não ser observados retrogenes autossómicos funcionais. Fizemos ainda uma análise evolutiva e funcional *in silico* dos ortólogos do retrogene *Rps4l*, inicialmente descrito em murganho, de modo a obter a idade aproximada da duplicação que originou o retrogene identificado em roedores.

Curiosamente, quer em roedores quer nos mamíferos analisados, as substituições aminoacídicas não toleradas em posições conservadas da RPS4 ocorreram sempre em *loops* nos domínios *N-terminal* e *RNA-binding*, cujas substituições são comuns também ao domínio *Central.* Os domínios *N-terminal* e *RNA-binding* estão ambos envolvidos em interações, mas no caso do primeiro domínio com o ribossoma e no caso do segundo com uma molécula de RNA.

1. Ortólogos da RPS4L em roedores

Da pesquisa por Blast com o retrogene *Rps4l* de murganho, selecionamos as sequências com maior percentagem de identidade com a sequência em questão e filogeneticamente relacionadas, potenciais ortólogos, o que resultou apenas em sequências de roedores.

A análise posicional e filogenética posterior confirmou a origem comum dos retrogenes da S4 encontrados e estabeleceu um limite inferior para o evento que originou a duplicação de 43,9 milhões de anos, tempo decorrido desde que *M. musculus* e *N. galili* partilharam um ancestral comum e superior de 82.8 milhões de anos, antes da radiação dos lagomorfos, já que não foi encontrado um ortólogo em *O. cuniculus*. No entanto, como não foram estudados outros lagomorfos, não podemos ter a certeza da origem deste evento, pois, poderão existir retrogenes não detetados devido ao facto de muitas das espécies deste grupo não terem ainda o genoma sequenciado. E assim, se se vier a confirmar a presença neste grupo, o evento da duplicação, será ainda mais antigo. Estudos anteriores apenas haviam detetado a RPS4L em rato e ratinho[38,60].

Após um evento de duplicação génica, as duas cópias podem ter trajetos evolutivos distintos, com pseudogenização ou aquisição de uma nova função por uma das sequências. Outra possibilidade é a subfuncionalização, em que o duplicado complementa a função do gene ancestral. Nos dois primeiros casos acumulam-se alterações na sequência da proteína, havendo um maior número de substituições nucleotídicas não sinónimas. No terceiro caso a complementaridade pode surgir por alteração da sequência proteica ou por alteração do padrão de expressão do gene. No caso dos retrogenes é frequente a aquisição de um padrão de expressão diferente do gene ancestral, dado que na maioria dos casos as sequências reguladoras do retrogene são as da região em que foi inserido no genoma após duplicação e diferentes das do gene ancestral[64]. A taxa de evolução dos retrogenes S4L identificados em roedores,

calculada através da razão dN/dS, indica uma baixa acumulação de substituições não sinónimas e portanto conservação da sequência ancestral S4X depois da duplicação e das S4L entre si, mesmo para o retrogene de *N. galili*, mais divergente. Em rato a RPS4L é expressa exclusivamente em testículo e em ratinho tem níveis de expressão mais elevados nesse órgão e é detetada ao longo de toda a espermatogénese[38,60]. O facto da taxa de evolução da RPS4L em murídeos ser baixa, com conservação da sequência ancestral e o padrão de expressão ser diferente da RPS4X poderá indicar subfuncionalização do retrogene, compensando a RPS4X durante a IMCS.

Fez-se ainda a análise funcional *in silico* dos ortólogos S4L aqui identificados em comparação com a RPS4X de *M. musculus*. Realizaram-se análises estruturais das proteínas mais semelhantes à RPS4X (todas exceto a de *N. galili*, mais divergente) de modo a avaliar o impacto das substituições aminoacídicas identificadas dentro dos domínios da proteína, em posições conservadas na RPS4 ao longo da evolução e consideradas deletérias por dois algoritmos (SIFT e PROVEAN Protein).

Os retrogenes de ***M. musculus*** e de ***R. norvegicus*** possuem apenas uma alteração aminoacídica não tolerada relativamente à S4X, que ocorre numa posição conservada (posição 95), em que uma treonina é substituída por uma serina, ambos aminoácidos hidroxílicos. O facto de os algoritmos (SIFT e PROVEAN Protein) classificarem esta substituição como não tolerada e deletéria dever-se-á ao facto de ser mais comum a presença da treonina relativamente à serina, nos alinhamentos feitos por ambos. Isto porque, em termos de propriedades, os aminoácidos são semelhantes. Ao nível das interações aminoacídicas, prevê-se que sejam mantidas inalteradas, mesmo após a substituição, e portanto, não deverá haver alteração da estrutura da proteína. A localização do retrogene encontrado em rato numa região sinténica relativamente à S4L do murganho (descrita por Hughes et al.[60] e confirmada pelo nosso estudo), a sua elevada semelhança com o ortólogo desta espécie (99,6%), o seu agrupamento filogenético com os ortólogos dos outros roedores (mais intimamente relacionado com o murganho) e a ausência de alterações a nível estrutural, indicam que o retrogene de rato é o ortólogo potencialmente funcional da S4L de murganho que é incorporada nos polissomas em testículo de ratinho[38].

Os retrogenes identificados em ***Cricetulus griseus, Microtus ochrogaster e Peromyscus maniculatus bairdii,*** apresentam a substituição descrita anteriormente em *R. norvegicus*, não apresentando outras substituições não toleradas dentro dos domínios, em posições conservadas da RPS4. Assim, a elevada semelhança dos

retrogenes encontrados com o ortólogo de murganho[38] - 99,2%, 99,2% e 98,9% para *C. griseus, M. ochrogaster e Peromyscus maniculatus bairdii,* respectivamente - o seu agrupamento filogenético com a S4L de murganho e rato, a ausência de alterações nas interações aminoacídicas bem como a nível estrutural, indicam que possamos ter identificado o ortólogo da S4L nestas espécies. Para a espécie *P. m. bairdii* a semelhança com a RPS4L é menor uma vez que esta sequência possui mais alterações aminoacídicas, relativamente às outras duas espécies, mas em posições não conservadas da RPS4 e portanto, não consideradas no estudo estrutural.

Mesocricetus auratus apresenta, além da substituição aminoacídica não tolerada descrita anteriormente em *R. norvegicus*, ainda uma outra na posição 163, em que um ácido aspártico é substituído por uma alanina. O ácido aspártico é um aminoácido polar, frequentemente encontrado em locais ativos ou de ligação na proteína. Contrariamente, a alanina é um aminoácido não polar, com uma cadeia lateral pouco reativa e por isso, raramente está envolvida na função da proteína mas pode ser importante no reconhecimento de substratos ou especificidade. No entanto, ambos pertencem ao grupo de aminoácidos pequenos e, apenas uma interação aminoacídica é afetada aquando da substituição, mantendo-se inalteradas as outras duas. Também não se verificam alterações estruturais quando se faz a sobreposição da RPS4X de murganho com o retrogene desta espécie, ainda que a baixa resolução da estrutura modelo disponível não permita determinar com precisão estas alterações. Assim, pensa-se que a substituição de um pelo outro, não causará danos de maior na função da proteína podendo ter sido classificada como deletéria por ambos os algoritmos, apenas pelo facto de não ser uma substituição frequente nas sequências dos alinhamentos realizados pelos mesmos. Deste modo, a elevada semelhança do retrogene com o ortólogo de murganho[38] (99,2%), o seu agrupamento filogenético com outros ortólogos, a ausência de alterações significativas nas interações aminoacídicas bem como a nível estrutural, indicam que possamos ter identificado a RPS4L nesta espécie.

Nannospalax galili apresenta a sequência mais divergente relativamente à S4X (85,6%), sendo também a mais divergente relativamente aos outros ortólogos analisados, com três substituições não toleradas e deletérias (SIFT e PROVEAN Protein) em posições conservadas. O retrogene apresenta ainda mais duas substituições aminoacídicas (posições 168 e 173) em que os algoritmos usados não estão de acordo quanto à classificação. Na posição 168 existe a substituição de uma lisina por uma arginina, ambas aminoácidos positivos e na posição 173 há a substituição de uma isoleucina por uma leucina, ambas aminoácidos alifáticos. Assim, pensa-se que

a substituição entre os aminoácidos envolvidos, na respetiva posição, não causará problemas a nível estrutural, o que vai ao encontro dos resultados apresentados pelo PROVEAN Protein. A divergência deste retrogene pode indicar que, devido ao facto do genoma desta espécie não estar ainda totalmente sequenciado e anotado, não foi identificado o ortólogo da RPS4L na nossa pesquisa mas sim, outra sequência mais divergente semelhante ao mesmo. No entanto, esta sequência sendo um duplicado da S4X poderá, também, ter sofrido um processo de especialização funcional, ainda que a razão dN/dS indique conservação da sequência aminoacídica. No entanto, dada a sua elevada divergência, a sequência não foi analisada estruturalmente.

2. Retrogenes da RPS4X em mamíferos

Dado que apenas se encontraram ortólogos da S4L em roedores, fizemos a pesquisa de homólogos da S4X por Blat em espécies de mamíferos selecionadas de modo a representar diferentes clados. Esta pesquisa resultou numa quantidade elevada de sequências não funcionais, pseudogenes, que acumularam ao longo da evolução algumas mutações (inserções, deleções de nucleótidos e/ou substituições) responsáveis pela ocorrência de modificações na grelha de leitura ou pela inserção de um codão de terminação prematuro. Assim, após uma cuidadosa anotação manual foram selecionadas as sequências homólogas da *RPS4X* com maior potencial para codificarem proteínas funcionais para análise evolutiva e funcional *in silico*.

A análise evolutiva destas sequências revelou vários eventos de retrotransposição ao longo da evolução dos mamíferos em todas as espécies analisadas, alguns muito recentes. A taxa de evolução dos retrogenes identificados, calculada através da razão dN/dS, indica uma baixa acumulação de substituições não sinónimas e portanto conservação da sequência ancestral S4X depois da duplicação, para as sequências encontradas no cromossoma 1 do coelho, 8 do cão e 23 da vaca. As restantes sequências mostram uma taxa de evolução mais elevada mas ainda bastante inferior a 1, à exceção da do cromossoma 9 da vaca, que parece estar a evoluir livre de pressões seletivas (aproximadamente igual a 1).

Realizou-se ainda a análise estrutural das proteínas codificadas por aqueles duplicados que apresentavam um número reduzido de alterações aminoacídicas previstas como

deletérias por dois algoritmos (SIFT e PROVEAN) e em posições conservadas na RPS4 ao longo da evolução, dentro dos domínios da proteína.

Orictolagus cuniculus

Em coelho identificamos dois retrogenes autossómicos potencialmente funcionais, nos cromossomas 1 e 14. Ambos possuem apenas uma substituição aminoacídica não tolerada (prevista pelo SIFT) em relação à S4X de murganho, que ocorre na posição 25, em que uma glicina é substituída por uma serina. A serina possui um grupo hidroxilo, muito reativo, o qual é capaz de formar ligações por pontes de hidrogénio com uma variedade de substratos polares. Contrariamente, a glicina, contém um átomo de hidrogénio na sua cadeia lateral, o que lhe confere maior flexibilidade conformacional, podendo ocupar partes de estruturas proteicas que são impossíveis para os outros aminoácidos. Pela sua particularidade, a glicina pode desempenhar um papel funcional distinto[65]. Esta substituição no domínio *N-terminal* da proteína, bastante conservado e que se pensa ser importante na interação com o ribossoma, poderá ter um impacto na função da proteína. No entanto para esta espécie encontra-se cristalizada uma proteína ribossomal (PDB: 4KZX)[58], a qual foi utilizada neste estudo para modelar a estrutura dos retrogenes encontrados em mamíferos que é idêntica ao retrogene do cromossoma 1. Portanto, tendo em conta a sua semelhança aminoacídica com a RPS4X - 98,5%, o facto de já ter sido cristalizada em associação com o ribossoma e a ausência de alterações estruturais significativas previstas pela sobreposição com a estrutura da S4X, este retrogene tem maior potencial funcional do que o retrogene do cromossoma 14 e poderá substituir a cópia da RPS4 ligada ao X durante a IMCS, na espermatogénese de coelho. No entanto, dadas as características descritas anteriormente para os aminoácidos substituídos e as diferentes interações previstas para a glicina e serina na posição 25, poderão haver modificações estruturais não detetáveis, as quais tenham implicações funcionais, ainda que a resolução da estrutura cristalográfica disponível utilizada para construir o modelo tridimensional da proteína codificada por estes retrogenes seja demasiado baixa (7.8 Å) para que estas previsões sejam precisas. Isto é válido também para o cromossoma 14, um pouco mais divergente da RPS4X (97,0% de similaridade). Portanto ainda que não seja possível assegurar que o retrogene do cromossoma 1 desempenha a mesma função que a cópia ligada ao X, sabe-se que a substituição aminoacídica observada não impede a incorporação no ribossoma da proteína codificada por este retrogene. Pelo exposto para a alteração aminoacídica comum entre os dois retrogenes, que parece não impedir a funcionalidade no retrogene 1, então também o retrogene 14 poderá ser funcional, salvaguardando que o mesmo

apresenta outras alterações aminoacídicas relativamente ao primeiro que podem influenciar a sua estrutura e consequentemente, a função.

Canis lupus familiaris

No cão foi perdida a cópia da RPS4 ligada ao Y[35] mas, através da nossa pesquisa, foi detetado um retrogene com potencial funcional no cromossoma 8. Este é muito semelhante à RPS4X, com uma semelhança aminoacídica de 98,9% e apenas uma alteração aminoacídica não tolerada na posição 32, onde há a substituição de uma serina por uma treonina. Segundo Betts e Russell[65], a serina pode ser substituída por outros aminoácidos polares ou pequenos, em particular a treonina que difere apenas por um grupo metilo, onde na serina está um grupo hidrogénio. Ambos os aminoácidos possuem um grupo hidroxilo muito reativo, tal como previamente descrito para a serina, e podem residir tanto dentro da proteína como à superfície. Segundo os nossos resultados, apenas uma das interações polares estabelecidas pela serina na RPS4X se manteve no duplicado, passando a treonina da posição 32 a interagir com a treonina da posição 33, com duas ligações. No entanto estas modificações não parecem ter consequências detetáveis a nível da estrutura global. Assim, dada a similaridade com a RPS4X, o facto de os aminoácidos substituídos pertencerem ao mesmo grupo e não haverem alterações estruturais globais visíveis, ainda que a resolução da estrutura modelo seja baixa (7.8 Å), este retrogene poderá ter mantido a função da RPS4X. O facto de tanto o SIFT como o PROVEAN terem previsto que esta substituição seria deletéria dever-se-á ao facto de ser mais comum a presença da serina relativamente à treonina nesta posição nas proteínas de diferentes espécies que estes programas utilizaram para avaliar a conservação desta posição já que as propriedades bioquímicas destes aminoácidos deverão permitir que possam permutar entre si. Apesar de recentemente ter sido descrito que a IMCS é incompleta nesta espécie[66] a *RPS4X* não foi identificada como um dos genes com aumento da expressão em espermatídeos e portanto deverá manter-se silenciada após a meiose. O retrogene do cromossoma 8 identificado poderá então ser um bom candidato à substituição da RPS4X nas fases finais da espermatogénese nesta espécie.

Bos taurus

Na vaca foram encontrados três retrogenes potencialmente funcionais através da nossa pesquisa, nos cromossomas 23, 9 e 7. Os retrogenes dos cromossomas 9 e 7 têm elevada similaridade relativamente à RPS4X (98,9% e 98,5%, respetivamente) e apenas apresentam alterações toleradas ou em posições não conservadas da RPS4. No entanto, são mais divergentes do que o retrogene do cromossoma 23, com 99,2% de semelhança e apenas uma alteração aminoacídica na posição 95, onde há a substituição de uma treonina na RPS4X por uma isoleucina no duplicado. Segundo Betts e Russell[65], a treonina pode ser substituída por outros aminoácidos polares, grupo ao qual não pertence a isoleucina. A treonina é comum em centros funcionais da proteína pois possui um grupo hidroxilo muito reativo, como descrito anteriormente para a serina, e contrariamente à isoleucina, que possui uma cadeia lateral não reativa e por isso raramente está envolvida em funções catalíticas, embora possa interferir no reconhecimento do substrato[65]. Segundo os nossos resultados, verificam-se alterações significativas nas interações polares para os aminoácidos alterados entre a RPS4X e o retrogene do cromossoma 23, sendo apenas mantida uma ligação comum aos dois aminoácidos susbtituídos. No entanto, não se verificam alterações estruturais globais visíveis, dentro dos limites que a resolução da estrutura permite. Durante o desenvolvimento deste estudo, foi publicado um trabalho em que se analisaram os retrogenes da RPS4 em vários mamíferos[60], no qual viram que o mesmo retrogene do cromossoma 23 da vaca (NCBI: XM_005223309.1) é expresso especificamente em testículo. Logo, a similaridade que este retrogene tem relativamente à S4 do cromossoma X, a ausência de modificações estruturais da única substituição aminoacídica potencialmente deletéria encontrada e o padrão de expressão sugerem que este retrogene será funcional e poderá ser importante na espermatogénese. Na vaca não existe cópia do gene S4 ligada ao cromossoma Y[9] e portanto a existência de um retrogene autossómico funcional deverá ter sido selecionada. Quanto aos outros dois retrogenes (cromossoma 9 e 7), não estão disponíveis padrões de expressão, mas nenhum apresenta alterações deletérias e portanto poderão ambos ser funcionais.

Equus caballus

No cavalo foi encontrado apenas um retrogene com potencial funcional, no cromossoma 6. Este retrogene é o mais divergente da respetiva S4X (92,5%) relativamente a todos os outros retrogenes estudados nas espécies analisadas. Como o cavalo não possui a cópia da RPS4 ligada ao Y, optámos por fazer também o seu estudo. Estão presentes cinco alterações aminoacídicas não toleradas, nas posições 31, 34, 44, 55 e 67. Para as posições 44 e 55, apesar da substituição, não estão presentes modificações nas interações aminoacídicas. Tal facto dever-se-á às propriedades químicas destes aminoácidos: a leucina e a valina são ambos aminoácidos alifáticos (posição 44) e, a alanina e a glicina pertencem ao grupo de aminoácidos pequenos (posição 55).[65] Assim, em ambas as posições, os aminoácidos substituídos podem permutar entre si. A arginina é um aminoácido polar e carregado positivamente, tolerando substituições por outros aminoácidos polares que é o caso da glutamina. Ambos são frequentemente encontrados em locais ativos e em sítios de ligação das proteínas, portanto não parece que esta alteração na posição 67 seja drástica a nível funcional[65]. A previsão de que estas alterações seriam não toleradas, atribuída pelo SIFT e pelo PROVEAN Protein, será devida à presença do aminoácido da RPS4X na maior parte das sequências do alinhamento. Na posição 31, deu-se a substituição de uma prolina por uma serina. A prolina pode em princípio ser substituída por outros aminoácidos pequenos, mas a sua estrutura complexa leva a que as substituições possam não ser bem toleradas e portanto, raramente é encontrada em locais ativos ou de ligação na proteína. Este aminoácido possui ainda uma cadeia lateral pouco reativa, contrariamente à serina que o substitui na posição 31 e que, por sua vez, dadas as suas propriedades já descritas, é frequente nos centros funcionais das proteínas[65]. Na posição 34, uma glicina é substituída por uma arginina que, como foi já descrito, pertencem a grupos muito diferentes, sendo que a sua alteração não deverá ser neutra. Então a classificação destas substituições como deletérias pelo PROVEAN Protein poderá refletir as diferenças nas propriedades dos aminoácidos. Também a nível estrutural, apesar de não se verificarem modificações no sítio onde ocorreram as substituições, observou-se que há uma *β-sheet* mais curta no retrogene relativamente à RPS4X. A *β-sheet* é formada por pontes de hidrogénio entre cadeias proteicas e, a ocorrência de substituições aminoacídicas em posições conservadas, pode levar à destabilização e desenrolamento da estrutura secundária[67]. Isto levará a alterações a nível terciário, o que se repercutirá funcionalmente. Assim, não podemos concluir que este retrogene mantém a função da RPS4X, mas poderá ser funcional e ter evoluído para desempenhar

outra função. Mais uma vez, o estudo dos padrões de expressão para este retrogene seriam essenciais para averiguar a sua especificidade e potencial funcional.

Para as espécies ***Felis catus*** e ***Homo sapiens***, não foram encontrados neste ou em outros estudos retrogenes autossómicos da *RPS4*, mas ambas as espécies possuem duas cópias no cromossoma Y[7,32,35]. Para as cópias do gato não foram descritos os padrões de expressão na literatura. Quanto à espécie ***Sus scrofa***, ainda não é certa a perda da cópia do gene *RPS4* ligado ao Y e, portanto, como não foram encontrados retrogenes autossómicos neste estudo, é provável que exista uma cópia ligada ao Y mas que ainda não foi detetada em estudos anteriores, já que nesta espécie o cromossoma Y não foi ainda totalmente sequenciado. Assim, o estudo dos padrões de expressão das RPS4Y em gato e porco é crucial para que melhor se compreenda a evolução e função destes duplicados nas diferentes linhagens de mamíferos.
A análise dos retrogenes autossómicos funcionais S4 de *M. musculus* e *R. norvegicus* (S4L) foi já descrita no capítulo anterior.

Um estudo recente propôs que a manutenção de retrogenes autossómicos da S4 seria potenciada pela perda da cópia do gene no cromossoma Y, para repor a dose da proteína em machos[60]. No entanto outros genes estão presentes apenas no cromossoma X de mamíferos e a compensação de dose entre sexos é atingida através da inativação destes genes num dos cromossomas X em cada célula das fêmeas. Assim a presença de retrogenes autossómicos da S4 com potencial funcional, em espécies que não mantiveram a cópia ligada ao cromossoma Y, poderá também ser explicada pela necessidade de repor a expressão do gene após a inativação meiótica dos cromossomas sexuais, na espermatogénese. Como os autossomas não são silenciados por este processo, os retrogenes autossómicos poderão ser expressos durante toda a espermatogénese e assim compensar o silenciamento dos cromossomas sexuais. Também as cópias ligadas ao cromossoma Y poderão ser expressas durante ou após a meiose e substituir o gene do cromossoma X. Não se pode também excluir a hipótese de que o fenómeno de IMCS seja incompleto em algumas espécies, tal como recentemente verificado no cão, e que o gene *RPS4X* seja expresso. O estudo mais refinado da região eucromática do cromossoma Y de coelho e porco, entre outros mamíferos, bem como os padrões de expressão dos genes e retrogenes da S4 nestas espécies irá contribuir para uma melhor compreensão da importância da IMCS na manutenção de retrogenes autossómicos.

Conclusão

Para os roedores analisados, os quais não apresentam cópia da S4 ligada ao Y mas apresentam potenciais ortólogos da *Rps4l*, não foram encontradas alterações estruturais que possam comprometer a função da proteína codificada pelos mesmos, o que juntamente com a baixa acumulação de substituições não sinónimas nos permite reforçar a hipótese de que poderão compensar a RPS4X nesta linhagem. A presença de ortólogos da S4l de ratinho nas espécies analisadas pertencentes a esta linhagem permite-nos datar o evento da duplicação, pelo menos, há mais de 43, 9 milhões de anos.

Em resumo, para o coelho identificamos pelo menos um retrogene funcional no cromossoma 1, que foi já detetado nos ribossomas em reticulócitos desta espécie[58]. Ainda que a sua função não esteja descrita, este tem potencial para compensar o gene do cromossoma X. Estudos futuros do cromossoma Y nesta espécie deverão confirmar a ausência de uma cópia do gene *RPS4Y*. Além disso seria interessante determinar o padrão de expressão do retrogene autossómico em tecidos de coelho para testar a hipótese de substituição da RPS4X na meiose. Na vaca, foi encontrado um retrogene funcional no cromossoma 23 que, pelo padrão de expressão específico em testículo, poderá substituir a RPS4X, já que a cópia ligada ao Y foi perdida nesta linhagem. No cão foi identificado um retrogene com potencial funcional mas cujo padrão de expressão é ainda desconhecido. Já o retrogene encontrado no cavalo apresenta substituições que poderão não ser bem toleradas e também não existem provas de que este retrogene será expresso. No entanto, dado que nesta espécie não existe cópia da S4 no cromossoma Y deverá existir um retrogene ativo, a não ser que a inativação meiótica seja incompleta no cavalo e que a RPS4X possa manter a expressão durante a espermatogénese. Finalmente, para o gato e o porco não foram encontradas cópias autossómicas, sendo que o gato tem duas cópias ligadas ao cromossoma Y, ainda que não seja conhecido o seu padrão de expressão.

Concluímos que, na maioria das espécies quando não existe uma cópia da S4 ligada ao cromossoma Y foram identificados retrogenes autossómicos da *RPS4X* com potencial funcional (murganho, rato e outros roedores, cão, vaca, coelho), os quais são candidatos a compensar a função da RPS4X aquando do seu silenciamento transcricional durante a IMCS na espermatogénese.

Referências bibliográficas

Literatura

1. Charlesworth, B. The evolution of sex chromosomes. *Science* **251,** 1030–1033 (1991).

2. Graves, J. A. M. Sex chromosome specialization and degeneration in mammals. *Cell* **124,** 901–914 (2006).

3. Smith, J. M. *The Evolution of Sex*. (Cambridge Univ. Press, 1978).

4. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508,** 488–93 (2014).

5. Ohno, S. *Sex Cromosomes and Sex-linked Genes*. (Springer, 1967).

6. Singh, N. & Petrov, D. Evolution of gene function on the X chromosome versus the autosomes. *Genome Dynamics* **3,** 101–118 (2007).

7. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).

8. Sinclair, A. H. *et al.* A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346,** 240–244 (1990).

9. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508,** 494–9 (2014).

10. Genetics Home Reference. (2015). at <http://ghr.nlm.nih.gov/gene/SRY>

11. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14,** 113–124 (2013).

12. Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278,** 675–680 (1997).

13. Cooke, H. J., Brown, W. R. & Rappold, G. A. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* **317,** 687–692 (1985).

14. Freije, D., Helms, C., Watson, M. S. & Donis-Keller, H. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science* **258,** 1784–1787 (1992).

15. Rice, W. R. Degeneration of a nonrecombining chromosome. *Science* **263,** 230–232 (1994).

16. Heard, E. & Disteche, C. M. Dosage compensation in mammals: Fine-tuning the expression of the X chromosome. *Genes Dev.* **20,** 1848–1867 (2006).

17. Lyon, M. F. Gene Action in the X-chromosome of the Mouse (Mus musculus L.). *Nature* **190,** 372–373 (1961).

18. Brockdorff, N. X-chromosome inactivation: Closing in on proteins that bind Xist RNA. *Trends Genet.* **18,** 352–358 (2002).

19. Charlesworth, B. The evolution of chromosomal sex determination and dosage compensation. *Curr. Biol.* **6,** 149–162 (1996).

20. Colot, V. & Rossignol, J. L. Eukaryotic dna methylation as an evolutionary device [review]. *Bioessays* **21,** 402–411 (1999).

21. Turner, J. M. A. Meiotic sex chromosome inactivation. *Development* **134,** 1823–1831 (2007).

22. McKee, B. D. & Handel, M. A. Sex chromosomes, recombination, and chromatin conformation. *Chromosoma* **102,** 71–80 (1993).

23. Turner, J. M. A. *et al.* BRCA1, Histone H2AX Phosphorylation, and Male Meiotic Sex Chromosome Inactivation. *Curr. Biol.* **14,** 2135–2142 (2004).

24. Xue, S. & Barna, M. Specialized ribosomes : a new frontier in gene regulation and organismal biology. *Nat. Rev. Mol. Cell Biol.* **13,** 355–369 (2012).

25. Lai, M. D. & Xu, J. Ribosomal proteins and colorectal cancer. *Curr Genomics* **8,** 43–49 (2007).

26. Alksne, L. E., Anthony, R. a, Liebman, S. W. & Warner, J. R. An accuracy center in the ribosome conserved over 2 billion years. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 9538–9541 (1993).

27. Nygård, O. & Nika, H. Identification by RNA-protein cross-linking of ribosomal proteins located at the interface between the small and the large subunits of mammalian ribosomes. *EMBO J.* **1,** 357–362 (1982).

28. Lake, J. A. Evolving Ribosome Structure: Domains in Archaebacteria, Eubacteria, Eocytes and Eukaryotes. *Annu. Rev. Biochem.* **54,** 507–530 (1985).

29. Zinn, a R., Alagappan, R. K., Brown, L. G., Wool, I. & Page, D. C. Structure and function of ribosomal protein S4 genes on the human and mouse sex chromosomes. *Mol. Cell. Biol.* **14,** 2485–2492 (1994).

30. Andrés, O. *et al.* RPS4Y gene family evolution in primates. *BMC Evol. Biol.* **8,** 142 (2008).

31. Jegalian, K. & Page, D. C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394,** 776–780 (1998).

32. Fisher, E. M. *et al.* Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell* **63,** 1205–1218 (1990).

33. Omoe, K. & Endo, A. Relationship between the monosomy X phenotype and Y-linked ribosomal protein S4 (Rps4) in several species of mammals: a molecular evolutionary analysis of Rps4 homologs. *Genomics* **31,** 44–50 (1996).

34. Pearks Wilkerson, A. J. *et al.* Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics* **92,** 329–338 (2008).

35. Li, G. *et al.* Comparative analysis of mammalian y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* **23,** 1486–1495 (2013).

36. Watanabe, M., Zinn, a R., Page, D. C. & Nishimoto, T. Functional equivalence of human X- and Y-encoded isoforms of ribosomal protein S4 consistent with a role in Turner syndrome. *Nat. Genet.* **4,** 268–271 (1993).

37. Lopes, A. M. *et al.* The human RPS4 paralogue on Yq11.223 encodes a structurally conserved ribosomal protein and is preferentially expressed during spermatogenesis. *BMC Mol. Biol.* **11,** 33 (2010).

38. Sugihara, Y. *et al.* Identification and expression of an autosomal paralogue of ribosomal protein S4, X-linked, in mice: Potential involvement of testis-specific ribosomal proteins in translation and spermatogenesis. *Gene* **521,** 91–99 (2013).

39. Uechi, T., Maeda, N., Tanaka, T. & Kenmochi, N. Functional second genes generated by retrotransposition of the X-linked ribosomal protein genes. *Nucleic Acids Res.* **30,** 5369–75 (2002).

40. Brosius, J. Retroposons--seeds of evolution. *Science (80-.).* **251,** 753–753 (1991).

41. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **18,** 292–298 (2003).

42. Ohno, S. *Evolution by gene duplication*. (Springer-Verlag, New York, USA, 1970).

43. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4,** 865–875 (2003).

44. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545 (1999).

45. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6,** 715–722 (1996).

46. Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. & Ravikesavan, R. Gene duplication as a major force in evolution. *J. Genet.* **92,** 155–61 (2013).

47. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).

48. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154,** 459–473 (2000).

49. Torrents, D. A Genome-Wide Survey of Human Pseudogenes. *Genome Res.* **13,** 2559–2567 (2003).

50. Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* **10,** 19–31 (2009).

51. Vinckenbosch, N., Dupanloup, I. & Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 3220–3225 (2006).

52. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).

53. Emerson, J. J., Kaessmann, H., Betrán, E. & Long, M. Extensive gene traffic on the mammalian X chromosome. *Science* **303,** 537–40 (2004).

54. Betrán, E., Thornton, K. & Long, M. Retroposed new genes out of the X in Drosophila. *Genome Res.* **12,** 1854–1859 (2002).

55. Kryazhimskiy, S. & Plotkin, J. B. The Population Genetics of dN/dS. *PLoS Genet.* **4,** e1000304 (2008).

56. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).

57. Drummond, A. J. *et al. Geneious v5.4*. *http://www.geneious.com* Available from www.geneious.com (2011). doi:http://www.geneious.com/

58. Lomakin, I. B. & Steitz, T. A. The initiation of mammalian protein synthesis and mRNA scanning mechanism. *Nature* **500,** 307–11 (2013).

59. DeLano, W. L. The PyMOL Molecular Graphics System, Version 1.1. *Schrödinger LLC* http://www.pymol.org (2002). doi:10.1038/hr.2014.17

60. Hughes, J. F., Skaletsky, H., Koutseva, N., Pyntikova, T. & Page, D. C. Sex chromosome-to-autosome transposition events counter Y-chromosome gene loss in mammals. *Genome Biol.* **16,** 104 (2015).

61. Paria, N. *et al.* A Gene Catalogue of the Euchromatic Male-Specific Region of the Horse Y Chromosome: Comparison with Human and Other Mammals. *PLoS One* **6,** e21374 (2011).

62. Bergen, A. W., Pratt, M., Mehlman, P. T. & Goldman, D. Evolution of RPS4Y. *Mol. Biol. Evol.* **15,** 1412–1419 (1998).

63. Hendriksen, P. J. *et al.* Postmeiotic transcription of X and Y chromosomal genes during spermatogenesis in the mouse. *Dev. Biol.* **170,** 730–733 (1995).

64. Ciomborowska, J., Rosikiewicz, W., Szklarczyk, D., Makałowski, W. & Makałowska, I. 'Orphan' retrogenes in the human genome. *Mol. Biol. Evol.* **30,** 384–396 (2013).

65. Betts, M. J. & Russell, R. B. in *Bioinformatics for Geneticists* 289–316 (John Wiley & Sons, Ltd). doi:10.1002/0470867302.ch14

66. Federici, F. *et al.* Incomplete meiotic sex chromosome inactivation in the domestic dog. *BMC Genomics* **16,** 291 (2015).

67. Browne, J. P., Strom, M., Martin, S. R. & Bayley, P. M. The role of beta-sheet interactions in domain stability, folding, and target recognition reactions of calmodulin. *Biochemistry* **36,** 9550–9561 (1997).

68. Blanga-Kanfi, S. *et al.* Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evol. Biol.* **9,** 71 (2009).

69. Graphodatsky, A. S., Trifonov, V. a & Stanyon, R. The genome diversity and karyotype evolution of mammals. *Mol. Cytogenet.* **4,** 22 (2011).

Recursos online

Ensembl: http://www.ensembl.org/index.html

fancyGENE: http://bio.ieo.eu/fancygene/

NCBI: http://www.ncbi.nlm.nih.gov/

NCBI BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi

PROVEAN Protein: http://provean.jcvi.org/seq_submit.php

PyMOL: http://pymol.org/academic

SIFT: http://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html

SWISS-MODEL: http://swissmodel.expasy.org/

TimeTree: http://www.timetree.org/

UCSC Genome Browser BLAT: http://genome.ucsc.edu/cgi-bin/hgBlat

Material suplementar

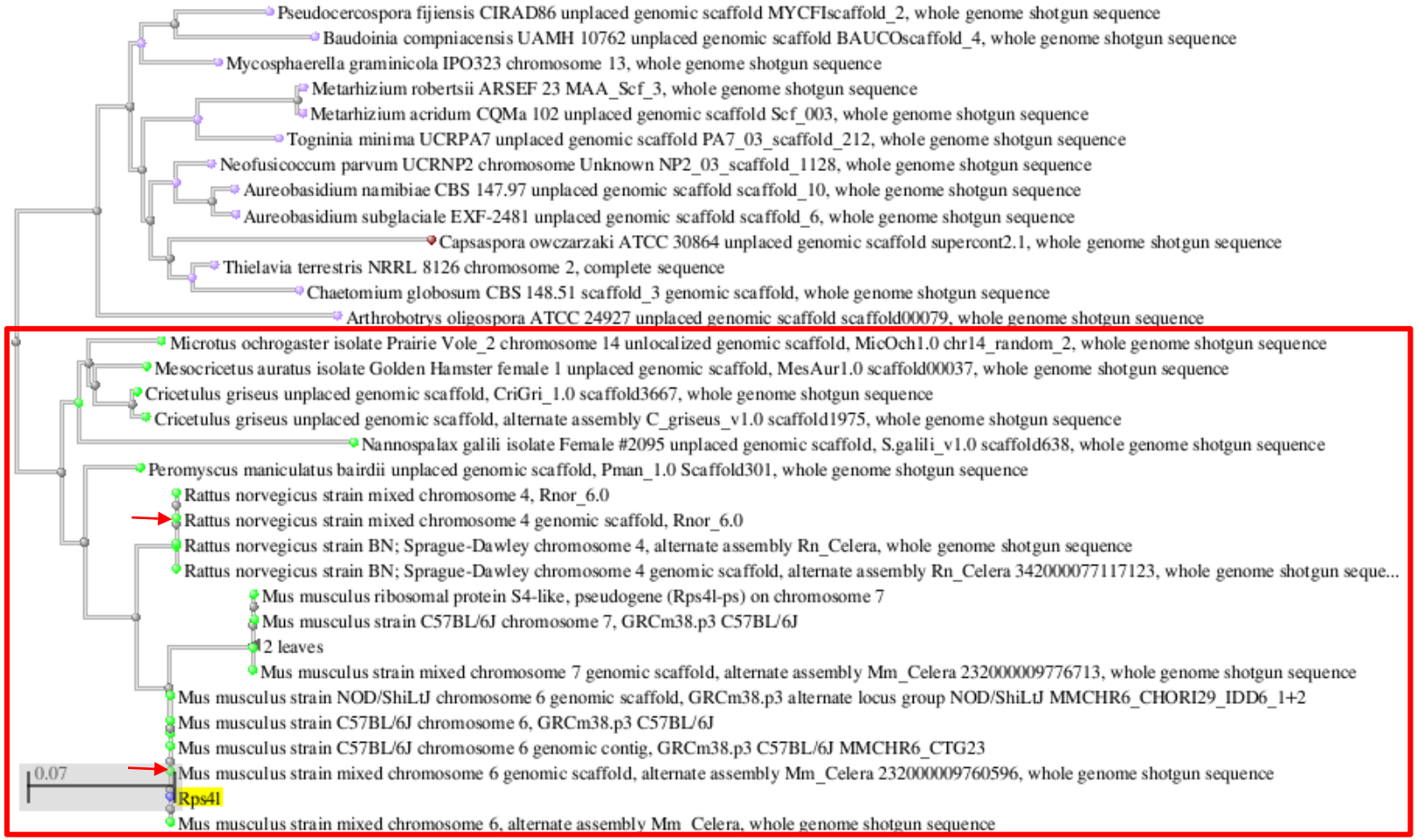


Figura A - Árvore de distância genética dos resultados do Blast com o CDS da *Rps4l*. As sequências mais próximas da *Rps4l* de murganho são as de outros roedores (dentro do retângulo vermelho). Para rato e murganho existem várias sequências resultantes de projetos de sequenciação alternativos, tendo-se selecionado para a análise as sequências assinaladas com seta vermelha.

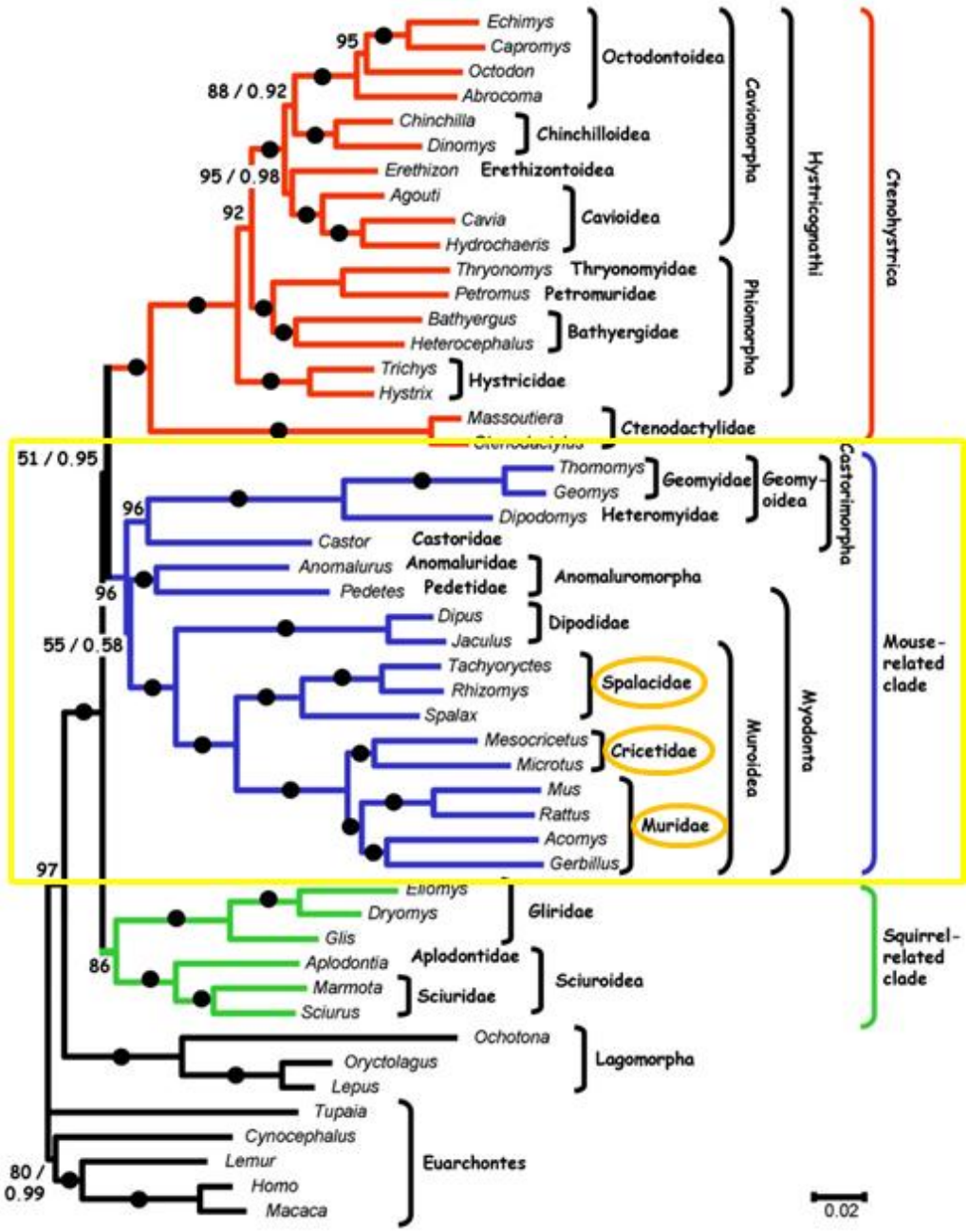


Figura B - Árvore filogenética dos roedores (adaptada de Blanga-Kanfi, Shani et al.[68]). No retângulo amarelo encontra-se destacado o clado relacionado com o murganho e com os círculos a laranja estão assinaladas as famílias às quais pertencem as espécies analisadas.

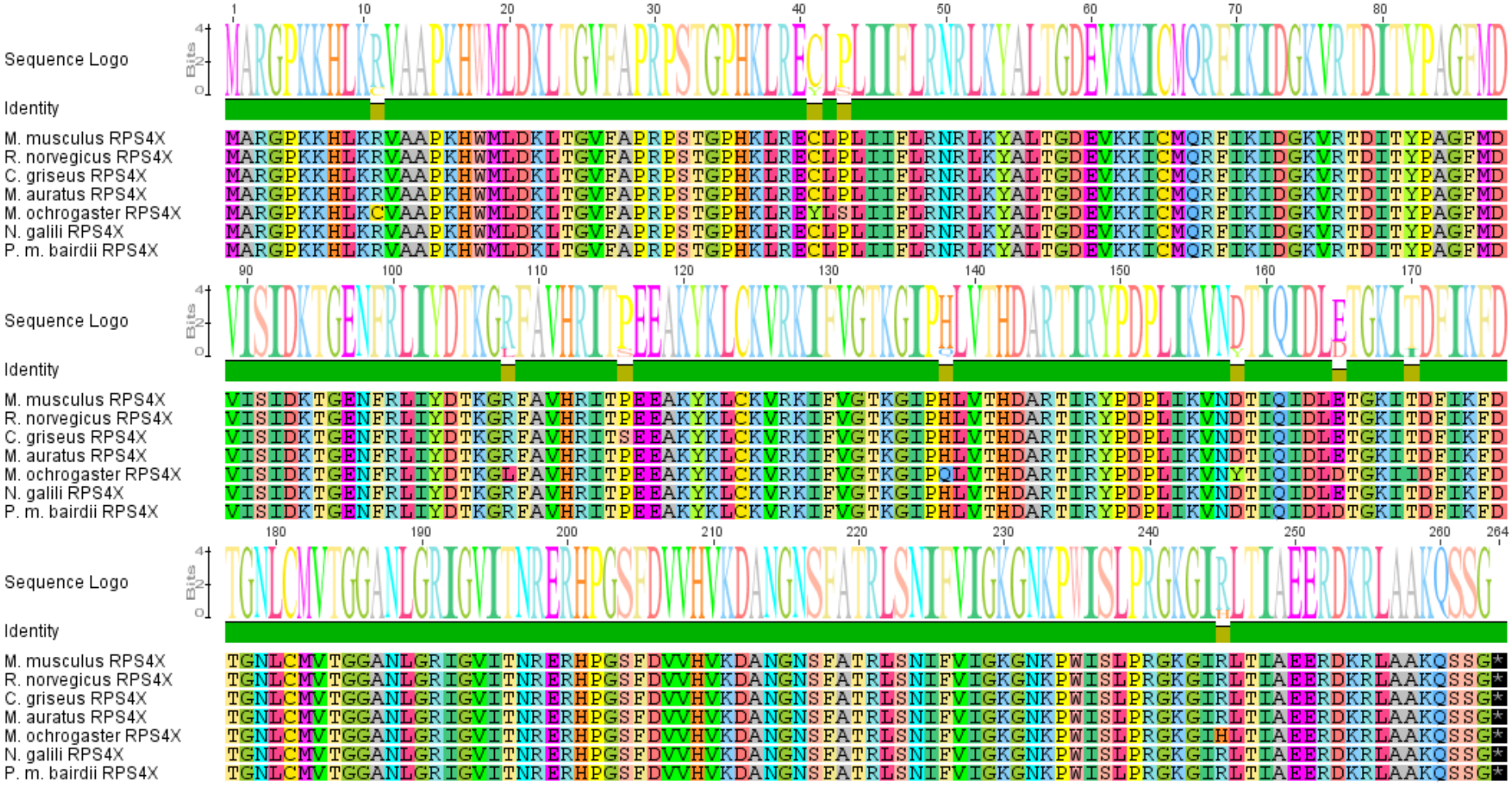


Figura C - Alinhamento das RPS4X das espécies de roedores analisados. Códigos de acesso das sequências: *M. musculus*: NP_033120.1; *R. norvegicus*: NP_001007601.1; *C. griseus*: NP_001233602.1; *M. auratus*: XP_005081267.1; *M. ochrogaster*: XP_005360594.1; *P. m. bairdii*: XP_006996111.1; *N. galili*: XP_008853778.1.

Análise da RPS4L de murganho

Tabela I - Posições não toleradas entre a RPS4X de murganho e a RPS4L.

Substituições não toleradas entre a RPS4X de murganho e a RPS4L					
Domínio	Posição	C/NC relativamente à RPS4	RP S4X	RP S4L	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	16	NC	K	R	N H R S K
	33	NC	T	A	i l r e d k v g Q n P A S T
	37	C	K	R	r P K
RNA-binding S4 (IPR002942)	45	NC	I	A	m f S G L V I
	69	NC	F	L	c w d p m e Q k g N r t s i a V H F L y
	72	NC	I	V	I V
	80	NC	I	V	c w m P d q g e n I r K v H T S a l f Y
	81	C	T	A	i f v l y p a H q e r T s K G d N
	95	C	T	S	T
	102	NC	I	V	A V I L
S4e, central region (IPR013845)	129	NC	I	V	c p m T K a l I V
	160	NC	I	V	L V I
	163	C	D	S	g E N S D
	165	NC	E	D	h m c i r n l q k V t P d g S E A
	166	NC	T	S	k d A N E S T
	172	NC	F	A	S Y F
Fora dos domínios	217	NC	S	G	m h i p l V g n r d Q A S T K E

*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão.

Análise dos duplicados da RPS4X de *Rattus norvegicus*

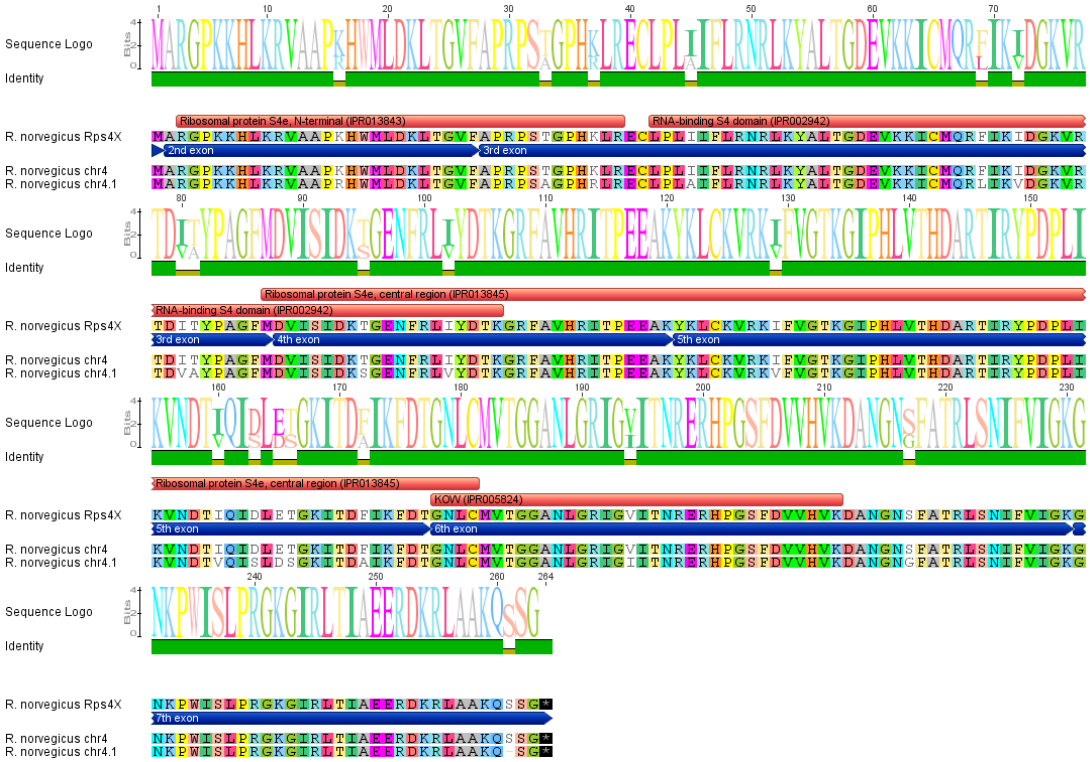


Figura D - Alinhamento da sequência proteica obtida por tradução *in silico* dos duplicados do rato selecionados para análise funcional. Ambos os duplicados estão localizados no cromossoma 4 que foram identificados como 4 e 4.1. O duplicado identificado como 4 tem uma similaridade aminoacídica de 100% com a RPS4X pelo que não faz sentido analisá-lo estruturalmente.

Análise dos duplicados da RPS4X de *Rattus norvegicus*

Tabela II - Posições não toleradas entre a RPS4X de rato e os seus duplicados (RPS4L).

Substituições não toleradas entre a RPS4X de rato e os seus duplicados					
Domínio	Posição	C/NC relativamente à RPS4	RPS4X	Chr4.1 (RPS4L)	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	16	NC	K	R	N H R S K
	33	NC	T	A	i l r e d k v g Q n P A S T
	37	C	K	R	r P K
RNA-binding S4 (IPR002942)	45	NC	I	A	m f S G L V I
	69	NC	F	L	c w d p m e Q k g N r t s i a V H F L y
	72	NC	I	V	I V
	80	NC	I	V	c w m P d q g e n I r K v H T S a l f Y
	81	C	T	A	i f v l y p a H q e r T s K G d N
	95	C	T	S	T
	102	NC	I	V	A V I L
S4e, central region (IPR013845)	129	NC	I	V	c p m T K a l I V
	160	NC	I	V	L V I
	163	C	D	S	g E N S D
	165	NC	E	D	h m c i r n l q k V t P d g S E A
	166	NC	T	S	k d A N E S T
	172	NC	F	A	S Y F
KOW (IPR005824)	194	NC	V	I	c q m p s e R K l A I T V
Fora dos domínios	217	NC	S	G	m h i p l V g n r d Q A S T K E

*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão.

Retrogenes da *RPS4X* em linhagens de mamíferos

Tabela III - Retrogenes da *RPS4X* encontrados por Blat nas espécies de interesse para este estudo.

Espécie	Cópias autossómicas	% de identidade nucleotídica	Bp
Mus musculus	Chr 2	98,7%	793
	Chr 13	98,1%	785
	Chr 5	97,8%	792
	Chr 7	97,0%	792
	Chr 18	96,6%	793
	Chr 3	95,2%	787
	Chr 13	94,2%	778
	Chr 1	90,9%	794
	Chr 7	84,4%	786
	Chr 6 (*Rps4l*)	80,3%	789
	Chr 7.1	78,8%	768
Rattus norvegicus	Chr 4	100%	792
	Chr 5	93,5%	780
	Chr 10	92,6%	751
	Chr 1	90,3%	779
	Chr 5	88,7%	764
	Chr 10	88,0%	752
	Chr 4.1	79,5%	789
	Chr 1	84,8%	761
Oryctolagus cuniculus	Chr 12	99,3%	792
	Chr 14	98,2%	792
	Chr 14	96,2%	777
	Chr 1	94,9%	792
	Chr 12.1	92,8%	791
	Un0028*	82,2%	766
Canis lupus familiaris	Chr 5	97,3%	792
	Chr 3	96,8%	792
	Chr 29	96,8%	792
	Chr 8	95,6%	792
	Chr4	95,5%	791
	Chr 20	94,7%	791
	Chr 3	89,8%	781
	Chr32	88,8%	780
	Chr 17	89,4%	769
	Chr7	87,8%	785
	Chr 12	87,2%	788
	Chr 31	87,1%	776
	Chr 22	85,0%	775
	Chr 22	84,8%	760
	Chr 36	85,4%	757
	Chr 13	86,1%	774
Felis catus	B1	89,1%	788
	A1	88.0%	786
	A1	88.0%	755
	A3	87,6%	783
	B2	87,1%	783
	D2	87,6%	779

	A1	85,1%	776
	C1	82,0%	782
	C1	84,4%	756
Bos taurus	Chr 9	99,0%	789
	Chr 7	98,9%	792
	Chr 3	97,8%	793
	Chr 20	96,7%	782
	Chr 8	97,0%	792
	Chr 10	96,9%	792
	Chr 11	96,8%	782
	Chr 4	94,5%	778
	Chr 23	94,2%	792
	Chr 11	91,6%	786
	Chr 16	89,2%	786
	Chr 21	88,6%	775
	Chr 1	88,3%	761
	Chr 13	89,9%	786
	Chr 16	85,8%	764
	Chr 2	83,6%	754
	Chr 7	82,1%	755
Sus scrofa	Chr 2	94,7%	798
	Chr 8	91,8%	790
Equus caballus	Chr 6	91,3%	792
	Chr 18	85,3%	792
	Chr 23	85,9%	789
Homo sapiens	Chr 5	94,7%	797
	Chr 6	93,8%	786
	Chr 13	93,4%	786
	Chr 11	94,0%	789
	Chr 20	95,3%	762
	Chr 12	93,5%	779
	Chr 10	93,2%	785
	Chr 12.1	92,7%	781
	Chr 6.1	91,6%	771
	Chr 17	92,1%	770
	Chr 19	90,6%	770
	Chr 2	89,5%	780
	Chr 6.2	90,2%	807
	Chr 1	88,3%	768

Selecionaram-se as sequências com base nos parâmetros (i) e (ii) descritos nos resultados. Está representada a sua localização cromossómica bem como a respetiva identidade nucleotídica com a *RPS4X* da mesma espécie e comprimento da sequência identificada. *Como não tem localização cromossómica, a sequência assinalada não foi analisada.

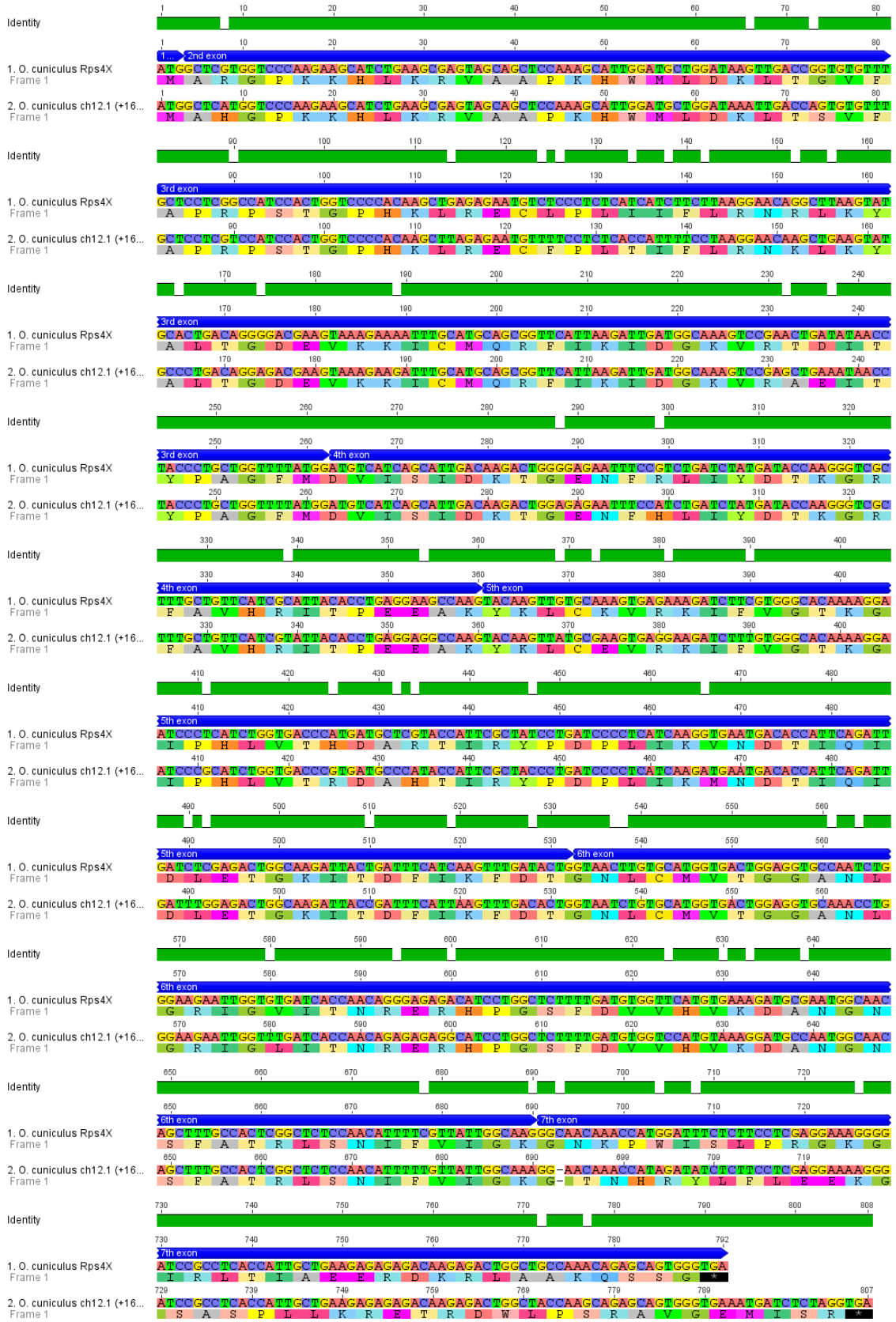


Figura E - Alinhamento e tradução *in silico* da sequência nucleotídica resultante do Blat da *RPS4X* de *O. cuniculus* no genoma da própria espécie. Devido a uma deleção *out of frame* de um nucleótido na posição 693 do alinhamento, esta sequência deixa de ter um codão STOP. Este é um exemplo em que se encontrou a sequência a 3' havendo um acréscimo de 16bp relativamente à sequência encontrada por Blat. A azul estão representados os exões anotados para a *RPS4X*.

Figura F - Alinhamento e tradução *in silico* da sequência nucleotídica resultante do Blat da *RPS4X* de *B. taurus* no genoma da própria espécie. A sequência apresentada tem menos 43 bp em relação à *RPS4X*, uma vez que apresenta um codão STOP prematuro devido à inserção de um nucleótido na posição 669 - adenina - que altera a grelha de leitura. A azul estão representados os exões anotados para a RPS4X.

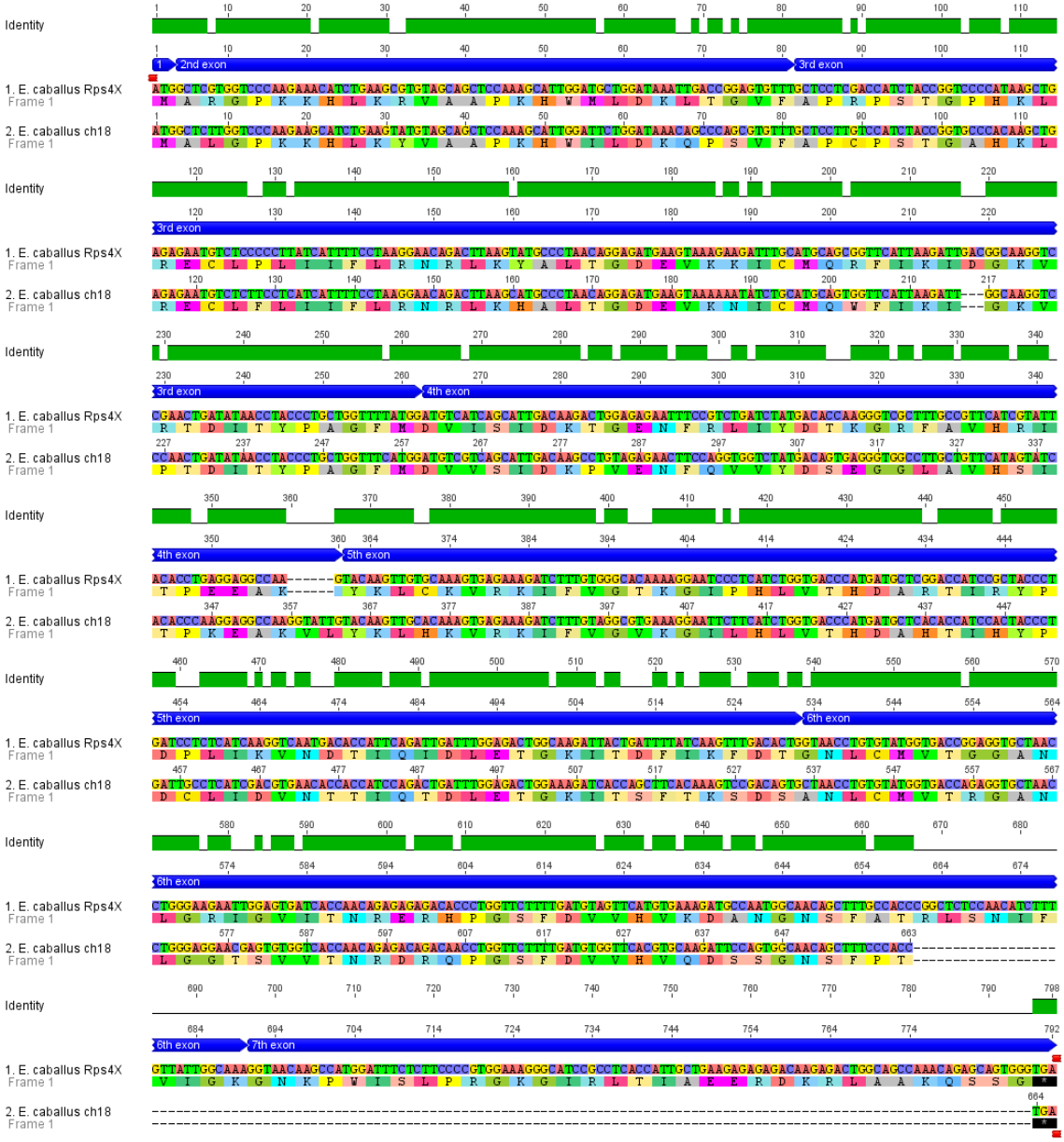


Figura G - Alinhamento e tradução *in silico* da sequência nucleotídica resultante do Blat da *RPS4X* de *E. caballus* no genoma da própria espécie. A sequência tem uma inserção de 6 nucleótidos e uma deleção de 126bp em relação à *RPS4X*, apresentando um codão STOP prematuramente. Esta sequência, apesar de mais curta do que os outros duplicados, apresenta ORF e mantém os domínios proteicos intactos. A azul estão representados os exões da *RPS4X.*

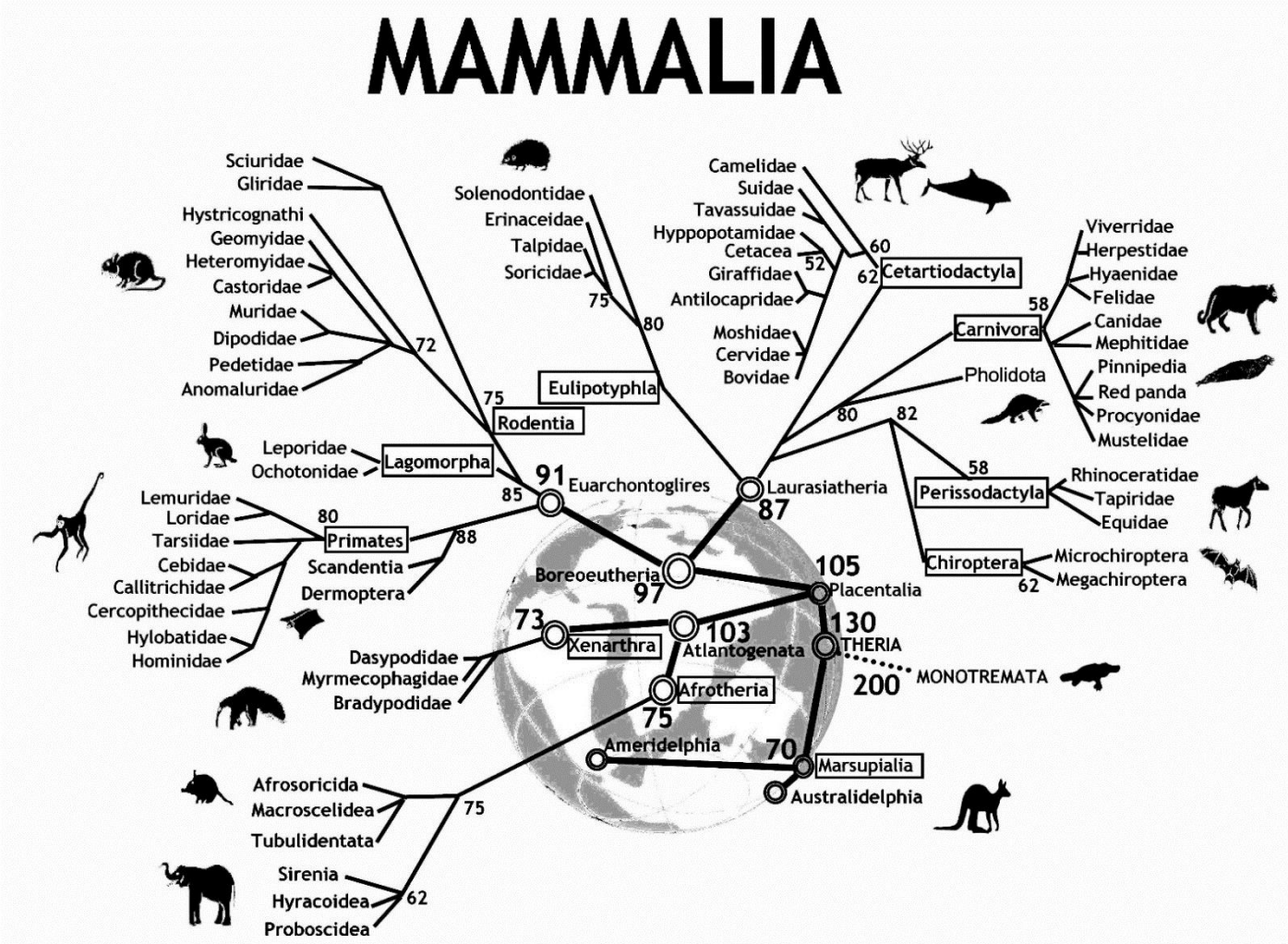


Figura H - Árvore evolutiva dos mamíferos. A árvore retrata as relações filogenéticas entre as ordens dos mamíferos. Os anéis duplos indicam superordens de mamíferos e os números indicam o tempo estimado das divergências. Adaptado de Graphodatsky et al.[69].

Análise dos duplicados da RPS4X de *Oryctolagus cuniculus*

Tabela IV - Posições não toleradas entre a RPS4X de *O. cuniculus* e os seus duplicados.

Substituições não toleradas entre a RPS4X do coelho e os seus duplicados								
Domínio	Posição[1]	C/NC relativamente à RPS4	RP S4X	Chr 1	Chr 12	Chr 14	Chr 12.1	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	3	C	R		C		H	T R
	9	C	L		P			L
	25	C	G	S		S	S	Y G
RNA-binding S4 (IPR002942)	45	NC	I				T	m f S G L V I
	79	C	D				E	s g C N D
	100	C	R				H	R
S4e, central region (IPR013845)	108	C	R		C			s h n a t l e F V q K R
	125	C	K				E	R K
	142	NC	H				R	l f a g n T D S Y H
	145	C	R				H	k L R
KOW (IPR005824)	208	NC	V			A		C L Y I V
Fora dos domínios	221	C	R		W	W		a k T S R
	234/233	C	K				H	Q E K
	237/236	C	I				L	V I
	238/237	C	S				F	E K T S
	241/239	C	P				E	P
	244/242	NC	K				S	e A D R K
	245/243	NC	G				A	G
	246/244	NC	K				S	V I
	247/245	NC	R				P	R K
	251/249	C	A				R	q f e k c S p m T L A I V
	253/251	NC	E				T	D K E
	261/256	NC	L				A	v R F M I L
	262/257	NC	A				V	Q K S G A
	264/259	NC	K				E	R Q K
	268/263	NC	G				R	G

[1]Posição do aminácido no alinhamento/Posição do aminoácido na sequência da RPS4X (dado o acréscimo de aminoácidos na sequência 12.1, houve necessidade de fazer a correspondência entre a posição dos aminoácidos no alinhamento com a sua respetiva posição na sequência para perceber as interações aminoacídicas na análise estrutural).
*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão.

Análise dos duplicados da RPS4X de *Canis lupus familiaris*

Tabela V - Posições não toleradas entre a RPS4X de *C. lupus familiaris* e os seus duplicados.

Substituições não toleradas entre a RPS4X do cão e os seus duplicados							
Domínios	Posição	C/NC relativamente à RPS4	RPS4X	Chr3	Chr8	Chr29	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	3	C	R	C		H	T R
	4	C	G			S	G
	8	C	H			Y	H
	9	C	L			V	L
	11	C	R	C			k T A R
	28	C	A	V			T A
	32	C	S		T		N A S
	35	C	P			H	P
RNA-binding S4 (IPR002942)	65	NC	C			W	C T M V I L
	77	C	R	Q			R
	100	C	R	H			R
S4e, central region (IPR013845)	142	NC	H			P	l f a g n T D S Y H
	145	C	R	H			k L R
	147	C	I			T	L I
	148	C	R	C		G	K R

*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado elo SIFT e as letras minúsculas resultam de uma previsão.

Análise dos duplicados da RPS4X de *Bos taurus*

Tabela VI - Posições não toleradas entre a RPS4X de *B. taurus* e os seus duplicados.

Substituições não toleradas entre a RPS4X da vaca e os seus duplicados										
Domínio	Posição [1]	C/NC relativa-mente à RPS4	RPS4X	Chr 23	Chr 9	Chr 7	Chr 8	Chr 10	Chr 3	Aminoáci-dos tolerados
s4e N-terminal (IPR013843)	3	C	R						Q	T R
	11	C	R				Y			k T A R
	22	C	K				E			Q R K
	30	NC	R						C	q M K R
	36	C	H						Y	H
RNA-binding S4 (IPR002942)	54	C	Y					C		Y
	70	NC	I				T			I F V I
	77	C	R				C			R
	92	NC	I			T				V L I
	95	C	T	I						T
	100/99	NC	R		H					Y F
	100	C	R				H			R
S4e, central region (IPR013845)	148	C	R						H	K R
	162	NC	I						T	F Y V L I
	173	C	I					M		A V I
KOW (IPR005824)	188	C	N				S	S		A N
	190	C	G						R	G
Fora dos domínios	220	C	T						K	s C T
	225	NC	I						H	C I V
	227	NC	V						R	a P L T I V
	228	C	I			T	F		Y	L I
	229	C	G						W	A G
	231	NC	G						R	e k Q T P S A G
	246/240	NC	R					H		n d q T A S R E K
	250/244	NC	I						Y	V I
	251/245	NC	R				C		S	R K
	253/247	NC	T						Y	T S
	254/248	C	I						H	k t R a E I V I

[1]Posição do aminácido no alinhamento/Posição do aminoácido na sequência da RPS4X (devido à alteração do quadro de leitura na sequência 3, houve necessidade de fazer a correspondência entre a posição dos aminoácidos no alinhamento com a sua respetiva posição na RPS4X para perceber as interações aminoacídicas na análise estrutural).
*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão.

Análise dos duplicados da RPS4X de *Equus caballus*

Tabela VII - Posições não toleradas entre a RPS4X de *E. caballus* e os seus duplicados.

Substituições não toleradas entre a RPS4X de cavalo e os seus duplicados						
Domínio	Posição [1]	C/NC relativamente à RPS4	RPS4X	Chr 6	Chr 18	Aminoácidos tolerados*
s4e N-terminal (IPR013843)	3	C	R		L	T R
	11	C	R		Y	k T A R
	24	NC	T		P	T E A S G
	25	C	G		S	Y G
	30	NC	R		C	q M K R
	31	C	P	S		P
	34	C	G	R		G
	35	C	P		A	P
RNA-binding S4 (IPR002942)	43	C	P		F	P
	44	C	L	V		L
	54	C	Y		H	Y
	55	C	A	G		A
	67	C	Q	R		E Q
	68	C	R		W	v p s h l n a t e Q G K R
	77	C	R		P	R
	95	C	T		P	T
	96	NC	G		V	s K D N G
	100	C	R		Q	R
S4e, central region (IPR013845)	108	C	R		G	s h n a t l e F V q K R
	117	C	E		K	D A E
	126/124	NC	C		H	m f y i q r d e n k v p L T S G C A
	147/145	C	R		H	k L R
	150/148	C	R		H	K R
	154/152	C	P		C	P
	164/162	NC	I		T	F Y V L I
	175/173	C	I		T	A V I
	177/175	C	F		S	a y v P m i F L
	180/178	C	G		A	N S D G
KOW (IPR005824)	193/191	C	R		G	R
	194/192	NC	I		T	I V
	195/193	C	G		S	G
	201/199	C	E		D	H Q E
Fora dos domínios	221/219	NC	A		P	t g V Q S A
	256/254	NC	K	V	-	h s n l t e q A R K

[1]Posição do aminácido no alinhamento/Posição do aminoácido na sequência da RPS4X (dado o acréscimo de aminoácidos na sequência 18, houve necessidade de fazer a correspondência entre a posição dos aminoácidos no alinhamento com a sua respetiva posição na RPS4X para perceber as interações aminoacídicas na análise estrutural).
*As letras maiúsculas indicam aminoácidos que aparecem no alinhamento realizado pelo SIFT e as letras minúsculas resultam de uma previsão.

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO